THESIS

Geographic profiling in biology

Mark Stevenson (Msc, BSc)

Thesis Submitted to Queen Mary University of London for the degree of Doctor of Philosophy July 2013.

Supervisors: Steve Le Comber, Rob Knell.

Funded by National Environment Research Council and Queen Mary University of London.

Abstract

In Chapter one I introduce the subject of geographic profiling, its use in criminology and its previous application to biology.

I go on in Chapter two to examine the original model and develop a likelihood-based approach to fit the parameters to data from 53 UK invasive species. GP performs well on this novel problem, and outperforms other simple spatial modelling techniques. Using simulations I show that GP is particularly efficient at locating sources when there is more than a single source.

Chapter three develops a Bayesian approach using Dirichlet Processes to account for the problem of multiple sources. This model was deloped in collaboration with Robert Verity. This new Bayesian model outperforms the original model used in criminology and offers a range of additional information from the data. The Bayesian GP model is then used to determine the sources of malaria outbreaks in Cairo. These developments significantly improve and extend the theory and application of GP.

In Chapter four I discuss the possible shapes of dispersal functions. I conduct a review of the literature and find a geometric mistake in the way linear distributions have been extracted from two-dimensional data. The correct back-transformation allows these dispersal distributions to be properly generated. Using this information; ecologists, conservationists and resources managers can now apply GP to real world problems and effectively allocate limited resources to locate sources of species invasions and disease outbreaks.

I go on in Chapter five to develop a method for fitting the primary parameter sigma

from the point pattern data and run simulations to show the effectiveness of this new approach.

In Chapter six I illustrate the application of GP to three problems, one in criminology, one in ecology and one in epidemiology. I finish by summarising the work in this thesis and discussing the potential future developments and applications of GP.

.

Acknowledgements

I would like to thank the invaluable and patient support of Steve Le Comber, who put up with my terrible timekeeping for three years. Steve's clarity, enthusiasm and attention to detail kept me on track throughout this project. Steve let me get on with my own work in my own time, but also knew when to rein me in and get things finished. I couldn't have done this without him.

I would also like to thank Kim Rossmo, for patiently explaining his work and contuining to engage with both Steve and myself as we applied his work to biological problems.

My thanks go to Annabel Dennis, Robyn Crowther, Jeyroy Gabriel and Carla Jackson. These project students who worked with me and GP throughout its development assisted by collecting and organising some of the data presented in this thesis.

I would also like to thank Robert Verity, for his help in getting me to understand Bayesian inference and for his amazing mathematical brain, as well as all of my labmates for providing a fun and engaging place to work.

I must also thank my parents for funding my degrees, which have led me to where I am now. They never stopped believing in me and supporting me, even though they won't understand a fraction of the work I now produce.

Finnaly the greatest help has been my beautiful and loving wife Xuan Xuan, who always thinks that I'm the most brilliant scientist in the world (even when I'm clearly not!) and made me the happiest man alive by marrying me even though I was still a struggling student.

‘Men honour what lies within the sphere of their knowledge, but do not realise how dependent they are on what lies beyond it.’

Zhuangzi

Contents

List of abbreviations

BRC	Biological Records Centre
CART	Categorical and Regression Tree
CGT	Criminal geographic targeting
CRP	Chinese restaurant process
DP	Dirichlet process
DPM	Dirichlet process mixture
EEA	European Environment Agency
ESBL	Extended-Spectrum Beta-Lactamases
EUNIS	European Nature Information System
GIS	Geographic information system
GP	Geographic profiling
H-PST	Hidden-Pick and Squash Tracking
MCMC	Markov Chain Monte Carlo
NBN	National Biodiversity Network
NERC	National Environment Research Council
NGO	Non-Governmental Organisation
QMUL	Queen Mary University of London

SDM	Species distribution model

List of figures

List of tables

Chapter 1: General introduction

This doctoral project seeks to apply geographic profiling models, first developed in criminology, to the population biology of invasive species and epidemiology. The project aims to develop these models, improve the mathematics of the models and fully test the resulting work with a range of data from different ecological problems.

1.1 Abstract

I describe in general terms the nature of animal movement and dispersal. I then outline the history of invasion biology and the methods used in this field, highlighting the similarity between the problems that face a criminologist and invasion manager. I introduce geographic profiling (GP), an approach first developed in environmental criminology to help solve serial (five or more) crimes. I go on to illustrate the history of the development of GP methodology as well as its applications. I describe in detail the mathematics of the criminal geographic targeting algorithm (CGT) first developed by Rossmo and show the outputs of this model. Finally I describe the progress in applying GP approaches to biological problems and summarise the research presented in this thesis.

1.2 Movement in biology

Organisms move in space and time, and understanding how they move is important in fields including population ecology, behavioural ecology, evolutionary biology, epidemiology, invasion biology and conservation (Kot *et al.* 1996; Clobert *et al.*

2012; Nathan 2001; Jonsen *et al.* 2003; Trakhenbrot *et al.* 2005). Animals interact with their environment in complex ways and these interactions can produce complex movement patterns (Jonsen *et al.* 2003). Understanding how these patterns arise and what their implications are for home-range and territorial dynamics (Moorcroft *et al.* 1999), climate change (Clobert 2004), habitat use and conservation (Belisle & St. Clair 2001, Block *et al.* 2001, Bowler & Benton 2005), biological invasions (Lewis & Kareiva 1993; Levin 2003), biological control (Jonsen *et al.* 2001), metapopulation dynamics (Moilanen & Hanski 1998) and community interactions (Ellner *et al.* 2001) are important issues in ecology.

Although the literature often uses the terms migration and dispersal interchangeably (Dingle 1996), in this thesis I will concentrate on dispersal. I will follow the definition of Ronce (2007) and define dispersal as 'the movement of individuals or propagules with consequences for gene flow across space'.

Given the importance of animal movement in all of the above fields it is not suprising that there has been a great deal of attention paid to modelling these processes; even so our ability to analyse movement patterns has been far outstripped by our ability to collect individual movement data (Jonsen *et al.* 2003). Here I will discuss some of the more common approaches.

The earliest approaches to modelling animal movment were based on random walks and their variants (for example correlated random walks and biased random walks) (Kareiva & Shigesada 1983, Turchin 1998, Sibert & Fournier 2001, Jonson *et al.* 2003). Unfortunately as Jonsen *et al*. (2003) point out, many of these use unrealistic assumptions (for example homogenous environments). Consequently recent interest has focussed on more complex models such as Lévy flight, where the distribution of

step lengths incorporates a heavy tail (Reynolds 2014).

Related to Lévy flight models (and random walks generally) are diffusion models (Magdziarz & Teuerle 2015). These models can be viewed as mathematically similar (Magdziarz & Teuerle 2015), but approach animal movement patterns from a different perspective. Diffusion models seek to simulate probabilistic patterns of movement across a population rather than by modelling individual movement patterns (Higgins & Richardson 1996). Reaction diffusion models were expanded by the work of Grünbaum (2000), whose work included interactions between the environment and organisms' internal states.

Kernel density estimation is a generalised non-parametric technique for estimating the probability density of a random variable (Rosenblatt 1959). Kernel density estimation has been used extensively, for example in estimating home range sizes (Worton 1989). However, being non parametric they cannot be used to drive hypotheis generation and make meaningful forecasts, instead just providing descriptions of movement patterns. Dispersal kenerls continue to be used (for example Lendrum *et al.* 2014), even though they provide an incomplete description of the dispersal process (Bowler & Benton 2005).

More recently state space models have found application in the study of animal movement. These models operate by using a series of observed (e.g. Markov models) and unobserved (e.g. Hidden Markov models) states to model a complex dynamic process, in this case movement (Jonsen *et al.* 2003).

As noted above, one of the areas where the study of animal movement is of major importance is the field of invasion biology. In the remainder of this chapter I first of all review different mathematical models used in invasion biology, before

introducing a new model based on an approach common in criminology but novel in biology.

1.3 Mathematical models used in invasion biology

Introduction

Invasive species are now viewed as the second most important driver of world biodiversity loss behind habitat destruction and have been identified as a significant component of global change (Vitousek *et al.* 1997; Wilcove *et al.* 1998; Wilson 1992). The cost of invasive species can run from millions to billions of dollars per occurrence (Jenkins 1996; Pimentel *et al.* 2001) as invasive species have been shown to affect native species through predation and competition, modify ecosystem functions as well as the abiotic environment and can spread pathogens (Strayer *et al.* 2006; Ricciardi 2007). It is for these reasons prevention and control of invasive species has been identified as a priority for conservation organisations and government wildlife and agriculture ministries globally (Hulme 2006).

Definitions

There have been a large number of definitions of invasive species. Many authors are particular to their own definition and the definitions often encompass slightly different concepts (Colautti & MacIsaac 2004). Terms and concepts crucial to understanding ecology have often been criticised for their tautological, ambiguous or

non-operational nature (McIntosh 1985; Peters 1991). The term 'invasion' was first used in relation to ecology by Goeze (1882) in his book 'Pflanzengeographie', in which he presented the invasion of the mango tree (*Magifera indica* L.) in Jamaica, which was viewed as a beneficial invasion. Other early authors such as Clements (1904) also used the term without any connotations of the negative or positive nature of the impact produced by the species in question.

The concept of an invasive species has now evolved to encompass a population-led approach, and invasive populations are considered important rather than species (Saki *et al.* 2001). The components of invasion are best understood when broken down into individual phases for assessment as part of an integrated framework that deals with each phase of the invasion in turn (Colautti & MacIsaac 2004).

History of the field

The issue of alien or exotic species, invading and establishing in areas where they have not been present historically, is a well-researched and well-developed field which had its origins in exotic species and island biogeography (Wallace 1880; Elton 1958; Baker & Stebbins 1965; Carlquist 1965; Carlton 1985; Crawley 1987; Drake *et al.* 1989; Hengeveld 1989; Williamson & Fitter 1996). Elton's (1958) book on invasions is a clear starting point from which invasion ecology began to emerge as a new discipline, yet Elton never truly defined the terms invasion or invader. This early literature attempted to answer what is now named the paradox of invasion (Elton 1958). This relates to fact that invasive species repeatedly invade and displace well-adapted native species despite having no prior selection for the novel environment (Elton 1958). The work has focussed on the characteristics that make

species good invaders. The key traits brought out have been life history traits such as *r* selection, large ranges, origins in large areas (continental) and preference for human disturbed habitat (May 1981).

The other main area of initial development was in the classification of habitats that made them more vulnerable to invasions (May 1981). The research primarily concluded that no strong individual traits could be associated with species in general that make them more likely to invade, but that repeatedly disturbed or human-altered habitat showed a greater prevalence of invading species (May 1981).

The field grew and received substantial global attention after the 1980s, with the publication of many symposia proceedings (Drake *et al.* 1989; di Castri *et al.* 1990; Groves & di Castri 1991; Pysek *et al.* 1995; Carey *et al.* 1996; Starfinger 1998; Mooney & Hobbs 2000) and reviews (eg Rejmanek *et al.* 2002). The field has seen developments both in theoretical understanding and in practical management tools. The field has now encompassed an understanding that the control and management of invasive species is an inherently interdisciplinary problem that involves the fields of ecology, economics and mathematics (Leung *et al.* 2002).

The framework of invasion

Following the work of numerous scientific areas such as weed scientists, resource managers, conservation biologists, restoration biologists, field ecologists and economists, a clear understanding of the stages that make up invasions and the relevant modelling and management steps that can be taken to prevent and manage invasive species (Sakai *et al.* 2001). Appendix A shows an example invasion framework (taken from Blackburn *et al.* 2011) that highlights the generalised steps in the invasion process and their relationship to management steps that could be taken. The movement, establishment and subsequent spread of invasive species is best characterised by a series of discrete steps, each of which poses different problems to both manager and modeller (Blackburn *et al.* 2011). Some of these stages are more relevant to prevention; others are more relevant for issues of control and restoration. There has been increasing understanding that feedback may occur between many of these steps (Sakai *et al.* 2001). Within each of the phases presented in Appendix A different types of predictive and analytical models can be used to gather information and make predictions on the risks presented by invasive species. Table 1.1 shows the main modelling methods used at each of the above stages of invasion. These models are then described in detail below.

Table 1.1 Invasion stages and different modelling approaches. The different stages of invasions are illustrated with the appropriate models used to infer information from these stages. For further details of the invasion stages see Blackburn *et al.* (2011).

Invasion stages	**Modelling approaches used**
Native elsewhere/Transport	Niche based modelling, spatial distribution models (SDMs), trait-based risk analysis
Introduction	Population dynamics models
Establishment	Evolutionary risk analysis
Spread	Gravity, wavefront and distance models
Ecological impact	Ecosystem models

Overview of relevant modelling approaches

The treatment of all modelling approaches used to detect, define and analyse the risks and damage caused by invasive species is beyond the scope of this paper. A summary of the methods of trait-based risk assessment, stochastic population growth models and niche modelling are presented as these most closely relate to or inform the new approach of geographic profiling.

Trait-based risk assessment

These models focus on the arrival prevention of invasive species and are based on the idea that invasive species have different biological traits to those that are not invasive (Keller *et al.* 2009). These ecological ideas have been formalised into a risk assessment structure to predict the impact of species before they are introduced (Keller *et al.* 2009). These models may be used to justify resource allocation and policy decisions aimed at preventing the arrival of identified high-risk species. Risk

analysis works by combining the cost/benefit trade-off of any decisions made with the understanding that scientific predictions are in themselves probabilistic (Orr 2003). The costs are thus multiplied with the conditionally independent future states and their probability of occurrence. This gives a distribution of future costs, which in turn is fed into a decision theory model that aims to maximise welfare over the given distribution of future costs (Keller *et al.* 2009). An example of a generic trait-based risk assement is shown in Table 1.2 below. The limitations of this approach include the fact that identifying reliable species traits that point to invasive species have repeatedly failed. A huge body of literature has attempted to identify those species that are invasive and demonstate that they tend to share certain traits (summarised by Pyšek *et al.* 2012), but few have ever reliably stood out over several studies. Goodwin *et al.* (1999) found that only species range was significant over a large study of 55 paired species groups and found it was predictive 70% of the time.

In addition, the validation of such models is difficult and has not received sufficient attention (Keller & Drake 2009). This method is normally achieved by splitting the data (if well resolved) into two sets, one for estimation and one for validation, based on the idea that the pool of introductions is constant and the introductions are independent (thus there is the idea that the traits are coming from an identically distributed distribution) (Keller *et al.* 2009). The model can also be tested with jacknifing or bootstrapping (Keller & Drake 2009).

Table 1.2 Generic non-indigenous aquatic organisms risk assessment (Orr 2003). Each factor is scored according to an expert opinion based on some form of data as being high, medium or low and scores are then aggregated to give a final risk assessment of high, medium or low.

1) Estimate probability of non-invasive being transported in a vector
2) Estimate individuals that will survive in a vector
3) Estimate probability of individuals becoming established on release
4) Estimate probability of population spread
5) Estimate magnitude of economic impacts
6) Estimate magnitude of environmental impacts
7) Estimate social/political issues

Another more sophisticated approach is the use of Categorical and Regression Tree (CART) analysis. CART (Kolar & Lodge 2002) has been used to predict if fish introduced to the Great Lakes will succeed or fail to establish. CART works by finding the split in predictor variables and maximising the within-group homogeneity of the two groups produced. It works like a reverse cluster analysis, slowly splitting the resulting groups to produce homogenous nodes. Data for each species is collected for each of the questions asked in turn. Of the 25 traits usually used in a risk assessment only four are needed when applied in a CART analysis (Kolar & Lodge 2002).

Stochastic population growth models

These methods assume that the species of interest has already been identified, by risk assessment or by data collected at a release site(s). These models seek to provide

quantifiable probabilities of the chances of establishment taking place. The probabilities created can then be used to evaluate economic costs and benefits of preventing invasions and for optimising management plans. Predicting ecological events is difficult, but there are clear statistical relationships between the propagule pressure (seeds or eggs or larvae etc) and the chance of establishment (Drake & Lodge 2004; Leung *et al.* 2004, Lockwood *et al.* 2005). The definition of establishment is something of a battleground (see Table 5.1 in Keller *et al.* 2009) but can be agreed upon as the term: 'When a species is unlikely to go extinct in the near term'. An example model taken from Haccau *et al.* (2005) takes the form of a dose response curve.

For quite a while now the growth and decline of colonising populations have been active areas of theoretical research (Crawley 1987). At the start of an invasion there will be have an arrival of a number of propagating individuals X_0, and the modeler seeks to establish the probability that the population will reach a certain undesired large size, conditional on establishment (May 1981). The size of the population at time t is give by X_t and it is required to be an integer. There are two important levels of uncertainty in this model. The first is that differing organisms will have different life histories (some will reproduce once, twice or many times etc) and hence we must construct a probability distribution to cover this uncertainty and then draw theoretical quantiles from this data (the vital rates) (Crawley 1987). The second source of variation is the choice of the model (flawed as always) in both structural uncertainty and parameter uncertainty (May 1981). The model chosen will be a discrete time model (for mathematical simplicity). Each organism is assigned a random number of offspring chosen from the distribution g(x); this is called the offspring distribution. ($g(0)>0$ and $0<g(0)+g(1)<1$) and assume no density

dependence. At time X_t then at X_{t+1} the number of individuals is the sum of X_t random draws from g(x): this is a discrete time Markov chain known as the Galton-Watson process (Galton & Watson 1874). There are two possible behaviors for this simple model. If the mean of g(x) is <=1 then the population will go extinct, but if it is >1 it will explode to infinity. The probability of extinction is given by the smallest non-negative root of the following equation:

$$f(z) = \sum_{x=0}^{\infty} g(x) z^x = z$$

(Equation 1.1)

f(z) is the probability generating function. This simple non-density-dependent function can be used to calculate the probability of a population explosion over a short time period such as an initial invasion. These models can be applied in the establishment phase, yet have been difficult to apply at the time of invasion owing to the strict demands they place on the investigator for information (Crawley 1987). It is necessary to collect large amounts of data, to attain reliable estimates for species vital rates (Crawley 1987).

Niche-based models

Habitat modelling has been the primary analytical tool used to illuminate the factors that can determine the chances of species successfully invading (Thuiller *et al.* 2005). The data on the distribution of invasive species can be used to correlate the environmental features determining the distribution and then build risk assessment maps (Thuiller *et al.* 2005; see also Elith & Leathwick 2006). Niche-based modeling operates under the nearly linear relationship obtained between certain key variables (such as latitude, temperature and geology) in species' native and potential introduced ranges (Wiens & Graham 2005). Essentially the correlation between environmental factors in the species' native habitat and environmental factors in the introduced habitat can be used to predict which species are likely to invade which area (Keller *et al.* 2009). This approach is then repeated with all the possible (or practical) environmental features and a logistic regression (or in some cases neural networks and other methods) used to pull together the correlations between areas (Keller *et al.* 2009). This gives the fundamental niche space of the invader in geographic space. The realised niche is of course influenced by the biological and historical realities, such as dispersal and competitors both intra- and inter-specific (Peterson 2003). In this respect niche models often fail to appreciate the path of invasion or key release sites such a ports and docks (in the case of aquatic invasive species).

Link to geographic profiling

One notable feature of all of these approaches is that the models typically run forwards in time; that is, they take the current locations of organisms, populations or

species and use these to predict their likely future locations. However, in many cases, it will be useful to employ models that run backwards in time – for example using the current locations of invasive populations to identify areas of introduction, or using disease case locations to infer sources of infectious disease (Le Comber *et al.* 2011). Such models are rare in biology, but common in criminology, where there are a number of approaches which aim to identify the home locations of serial offenders based on crime site locations. The most widely used of these is geographic profiling (Rossmo 2000).

1.3 Geographic profiling

History in criminology

GP did not begin as a mathematical/analytic tool; rather it grew out of the theories present in environmental criminology from an empirical basis into a practical tool to assist on-going investigations (Rossmo 2000). The history and development of the approach is important as it allows understanding of the useful components that could be applied to wider problems in ecology and can also highlight the components that are specific to a criminal investigation.

The key components that coalesced together to make GP possible were (i) the theory of environmental criminology created by Brantingham & Brantingham (1981); (ii) the mathematical modelling of distance decay functions by Capone & Nichols (1976); (iii) the incorporation of point pattern analysis first developed in ecology, and (iv) the growth in computer technology and geographic information systems (GIS).

The study of spatial crime patterns has a long history. Work on human sociology often had an explicitly spatial component (Burgess *et al.* 1921). The concentric zone model developed by Ernst Burgess in 1924 was an early approach to describe the zoning of human social groupings within urban areas (Brantingham & Brantingham 1981). These many different approaches did not fall under one clearly defined research directive until the work of Brantingham & Brantingham (1981) created the field of environmental criminology. This field focuses on the study of all aspects of crime (criminals and victims, etc) based on rational offender choices (Harrow *et al.* 1993), when they are considered in the light of particular locations, as well as the movement and activity patterns of individuals of groups and how these influence criminality (Harries 1990).

This approach was developed in the 1980s by the married partnership of Paul and Patricia Brantingham. The novelty of their method was to place the focus of research into criminology on the locational and contextual factors that influence criminality, rather than on the offenders themselves. The key components of the approach are space, time, offenders, targets, law/deterrents and geography (Brantingham & Brantingham 1981). For a crime to occur, space, time, an offender, a target and some aspect of law are all needed. These five are described as the minimum possible set of environmental conditions required to constitute a crime (Brantingham & Brantingham 1991). This field placed the emphasis on data such as land usage, street design, traffic patterns, public transport schedules, daily activities and routine movements of both criminals and victims.

Environmental criminology as codified by Brantingham and Brantingham had an explicit spatial dimension. They created an organised system for assessing a criminal's activity space known as micro spatial analysis of crime. This approach

looked at the pattern of travel that offenders make, as well as their awareness of their environment as they move through it. This framework allowed the estimation of the activity and awareness space of common offence types based on offender location. This activity space approach laid the groundwork for GP to be developed. It was in this theoretical framework that Kim Rossmo, who developed GP, worked as a doctoral candidate under the Brantinghams.

The modelling of movement decay functions of commuters, shopper and criminals was developed by Capone & Nichols (1976). They were the first authors to fit a range of competing models to the distances that criminals moved to commit crimes. They found that three functions all worked well: the Pareto function, the exponential and the combined Pareto exponential. All three have an exponential shape, with many close occurrences falling off so that there is a long tail in which a few events occur (Capone & Nichols, 1976). Rossmo initially used a combined Pareto exponential in his model fitting before moving to an exponential (Rossmo, 2000).

Whilst this approach was being developed other authors incorporated approaches developed in population ecology for the analysis of point pattern data. Smith (1975) wrote at length about the use of nearest neighbour analysis of point pattern data. Originally developed in ecology, this approach looks at the distances between points inferring the randomness or regularity of the data. This approach informed Rossmo (2000) who used a similar method to extract the within-cluster distances in order to fit his GP model.

The development of efficient computer mapping tools was a great boon to criminologists. Pin maps had been used since policing had been formalised in England in the 18th century. The use of GIS technology allowed both the mapping of

point data and the uploading of district boundaries, housing price zones, street lighting maps and police presence. Mapping increased in importance and complexity with the use of 3D contour maps and wireframes.

The primary tools before the development of GP were maps and other simple spatial statistics. The field has recently expanded to also include other approaches, including hot spot analysis, metric topology, kernel density models and other approaches (Newton & Newton 1985; Verma & Lodha 2002).

The Rossmo approach

Drawing on the routine activity theory of Brantingham & Brantingham (1982), Rossmo made an important logical jump. Instead of looking at the offender's movement to try to predict crimes, he used crime locations to try to predict the location of the offender's home. This use of inverse logic to infer the location of the offender from crimes is very similar to that undertaken by inverse probability (see Chapter 3). Rossmo did not see or phrase his approach in these terms, but nevertheless the approach he uses bears striking similarity to a Bayesian one.

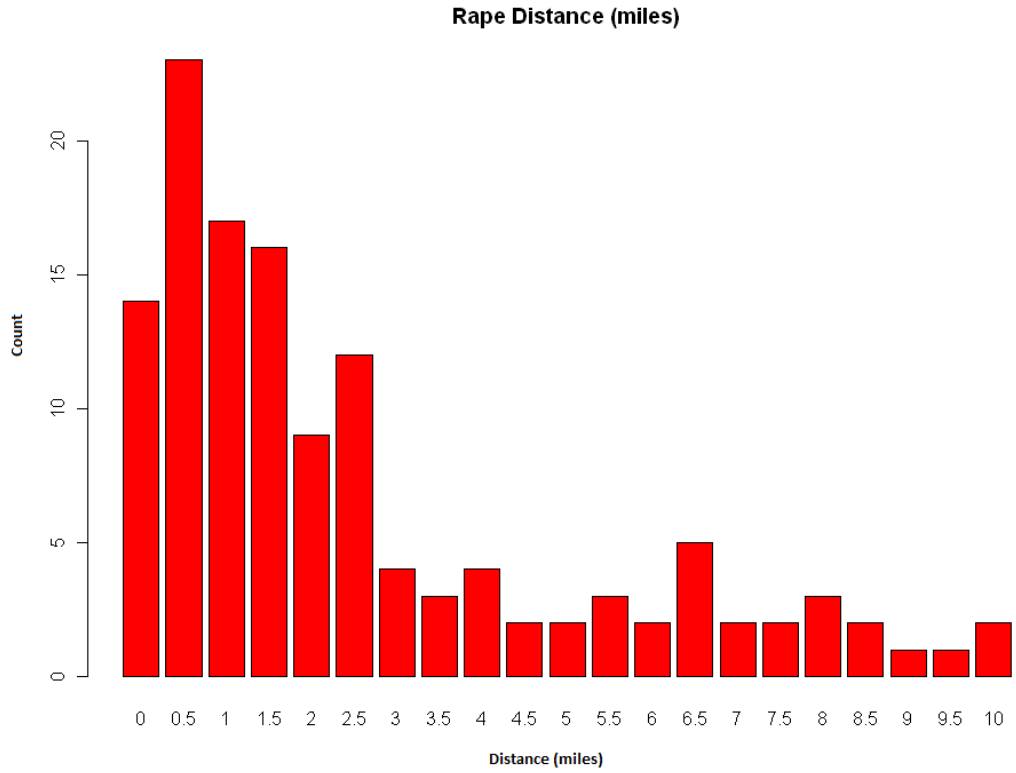


Figure 1.1 Distribution of known offender distances. The figure, redrawn from data presented in Rossmo et al. (2004), shows the frequency of rapes from different offenders on the y axis plotted against distance in miles on the x axis. This data shows the characetristic Rossmo function shape, coming to a peak after a buffer zone and then falling off.

Using set theory and overlapping sets, Rossmo first describes the overall problem. He then notes that these distributions of movement are not simple sets, but rather distributions drawn from certain functions. He describes the decay function as both a negative exponential and Pareto function. This is taken from the work by Capone & Nichols (1976) in which distances of shopping trips and others were described using a range of functions. They eventually settled on a combined Pareto and exponential function that Rossmo adopted. He also used data from known crime trips to help inform his shape of distribution (Rossmo 2000). Known offender distances of rapists

and a collection of several types of crimes are shown in Figures 1.1 and 1.2 (Hazelwood, 1987; LeBeau 1992; Sapp 1994).

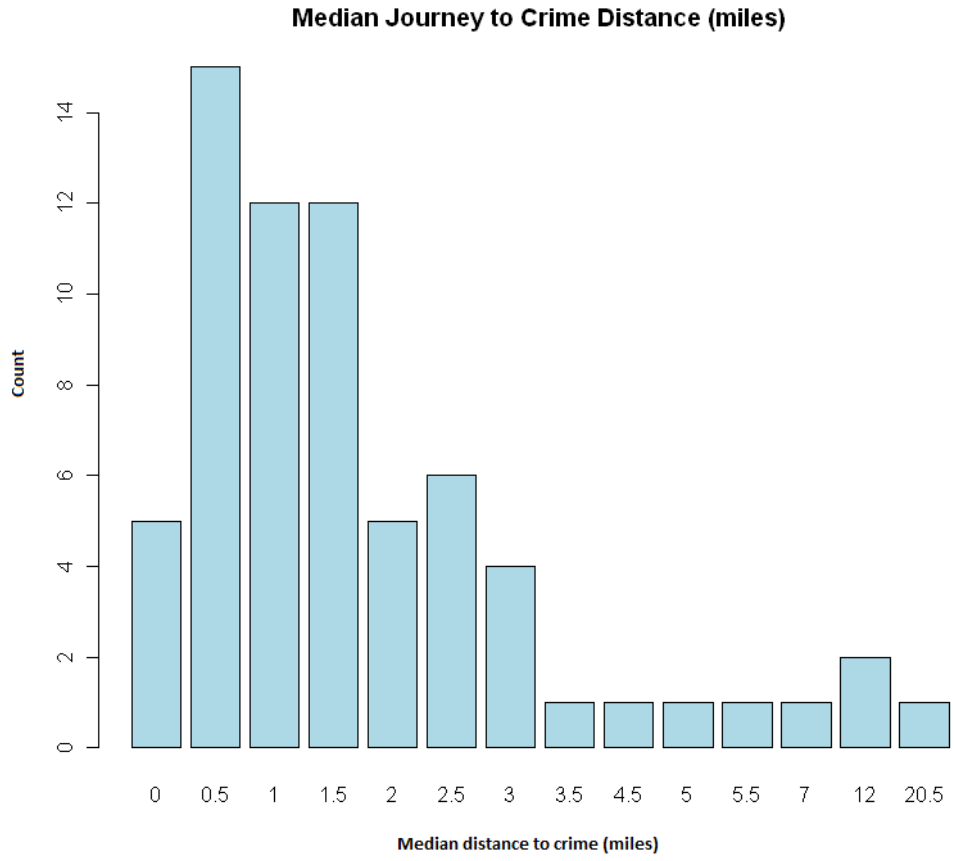


Figure 1.2 Distribution of journey to crime distances, redrawn from data presented in Rossmo *et al.* (2004). The figure shows the median frequency of a range of crimes from different offenders on the y axis plotted against distance in miles on the x axis, again showing the characteristic Rossmo function shape.

From the shape of these distributions Rossmo arrived at the idea of a buffer zone. This idea has been contentious in criminology since its formation. Yet there is strong evidence for it in a selection of offences such as those shown in Figures 1.1 and 1.2. Rossmo combined the two ideas of negative exponential/Pareto distance decay and

the buffer zone create a new distribution used in his criminal geographic targeting (CGT) algorithm (Rossmo 2000). This Rossmo distribution takes the shape of a two-part curve that when seen in three dimensions appears much like the caldera of a volcano. An example of the Rossmo distribution is shown in Figure 1.3.

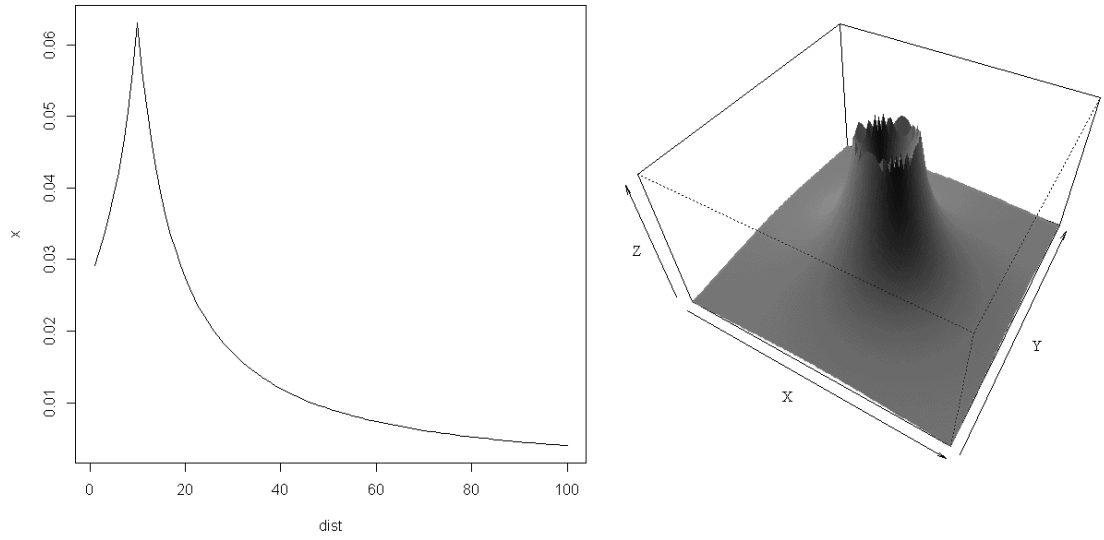


Figure 1.3 A simple diagrammatic representation of the Rossmo distribution. This Rossmo distribution (see Chapter 2) takes the shape of a two-part curve, first rising to a point at the radius of the buffer zone before falling off with distance. Left: the Rossmo distribution in two dimensions. Right: the same function, in three dimensions.

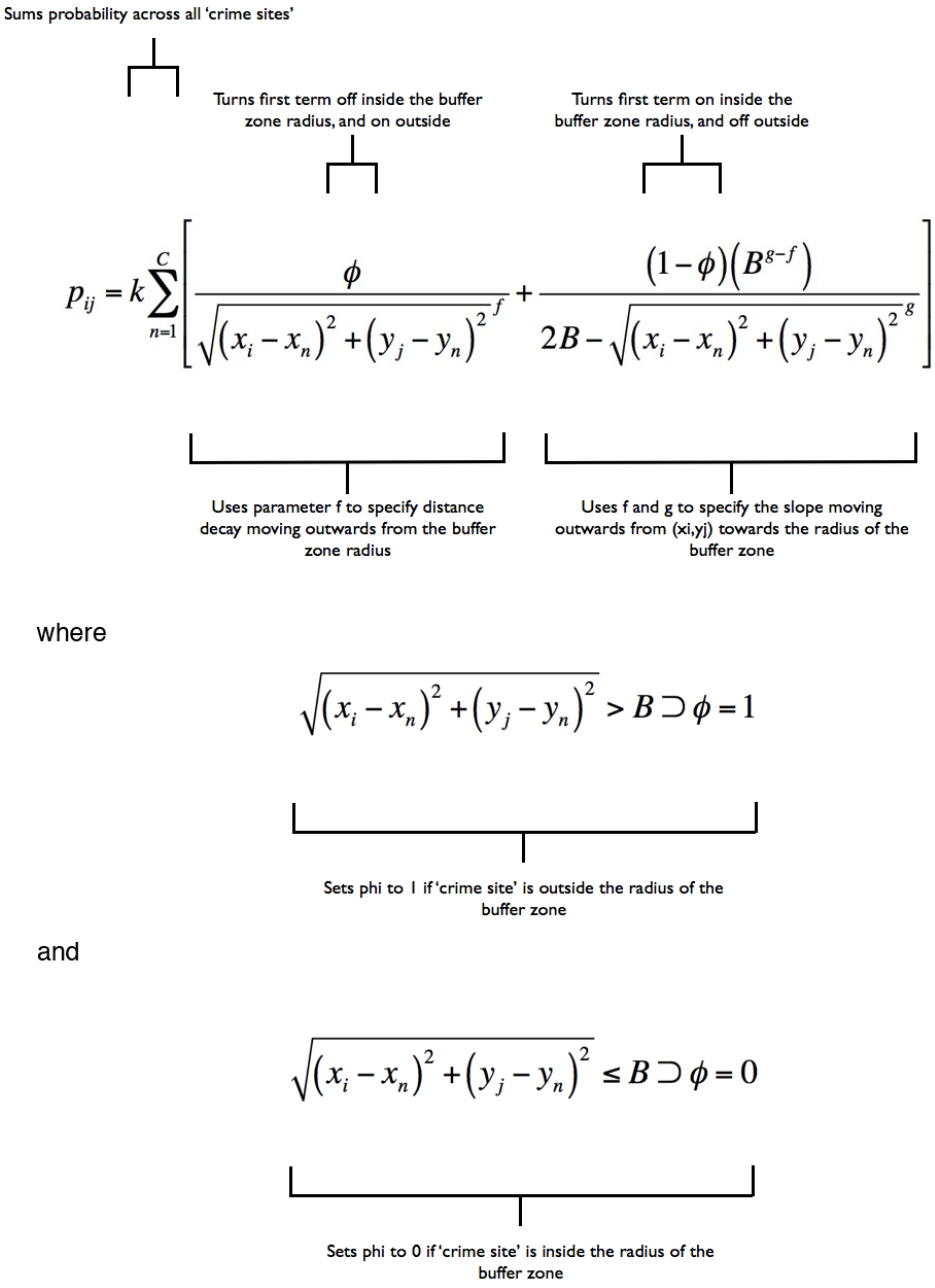
A full description of the mathematics of the Rossmo model is provided in Section 1.5. Since GP's initial development by Rossmo it has been shown to be a method that is easily applied to a range of problems. While it was originally applied to the cases of serial murder (Kind 1987; Dorney 1990), it was also used for investigations involving rape, sexual assault, arson, robbery, bombing, kidnapping, burglary, auto theft, fraud, vandalism, and graffiti (Rossmo 2012). This led to the development of a

program for training crime analysts, researchers and law enforcement officers in the techniques of GP. The Geographic Profiling Analysis (GPA) course grew out of part of a technology demonstration program first introduced in 2001 (Laverty & MacLaren 2002). Now more than 600 people from 14 nations (including the author) have been trained in the use of this technique.

GP has turned out to be a remarkably robust and applicable methodology (Rossmo & Harries 2011). It has expanded beyond its original application to range of new areas and continues to do so. It has also recently been used for border security (Rossmo & Velarde 2008), military counterinsurgency (Brown *et al.* 2005) and counterterrorism (Rossmo & Harries 2011). These techniques have similar goals and problems to the original application in criminology, in which prioritising a list of suspects to best direct limited investigative resources using spatial information to support existing evidence is a pressing concern. GP has a demonstrated ability to assist in these and other similar types of problems (Rossmo & Harries 2011).

1.5 Description of the Rossmo model

A full description of the model can be found in Rossmo (2000). Here, I describe a slight variant, introduced first by Le Comber *et al.* (2006), which uses Euclidean rather than Manhattan distances (this approach was chosen as there is no reason that biological species should move in the restricted pattern defined by urban (particularly North American) street layouts). The geographic profiling function generates a prioritised surface that describes the optimal search pattern for the sources of invasive species (Figure 1.4).



where

and

Figure 1.4 An annotated version of the Criminal Geographic Targeting (CGT) algorithm from Rossmo (2000). For each point (i,j) within the target area, the score function (p) is calculated as shown, such that ϕ functions as a switch that is set to 0 for sites within the buffer zone, and 1 for sites outside the buffer zone. *k* is an empirically determined constant, *B* is the radius of the buffer zone, *C* is the number of sightings of the organism, *f* and *g* are parameters that control the shape of the decay function, (x_i,y_j) are the coordinates of point (i,j) and (x_n,y_n) are the coordinates of the nth site. Thus, p_{ij} describes the likelihood that the anchor point occurs at point (i,j), given the crime site locations

(Rossmo 2000). The equation describes a two-part curve, which when plotted in three dimensions resembles the caldera of a volcano. When summed these 'volcano' shaped decay functions produce a surface that describes an efficient search pattern for the location of criminal anchor points.

Some example surfaces produced by the Rossmo function are shown in Figure 1.5. As can be seen from both the history of the model described in Section 1.3 and its technical description above, this model is very different to the existing approaches being used in spatial ecology.

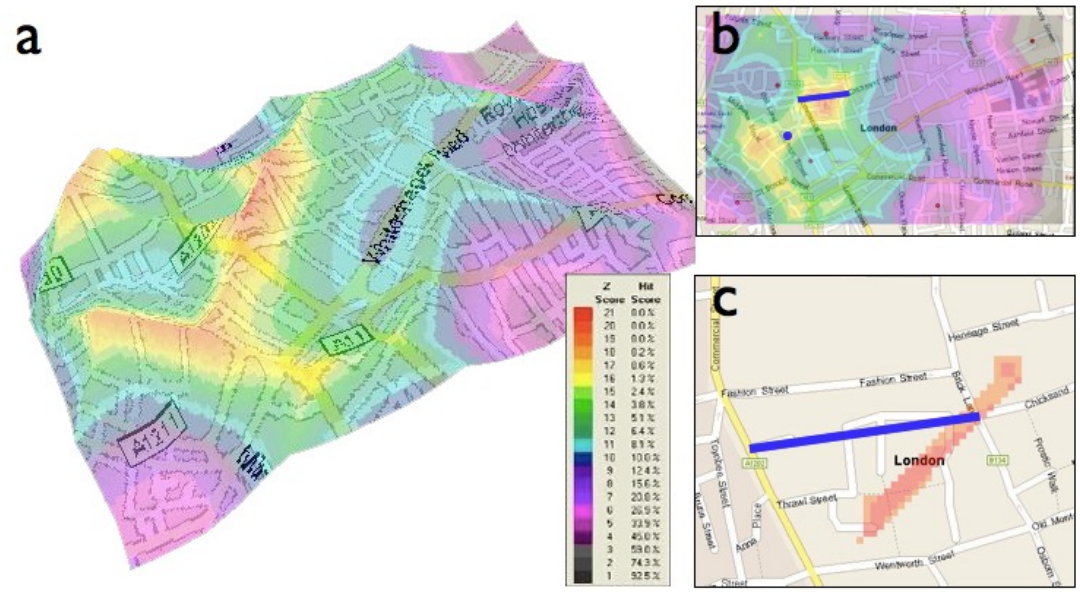


Figure 1.5 The Rossmo function applied to the Jack the Ripper case (a) The geoprofile produced by the Rossmo function from analysis of the body dump sites of the five canonical murders attributed to Jack the Ripper in the East End of London in 1888; higher areas of the geoprofile are shown in red, and lower areas in blue (b) a two-dimensional version of the same geoprofile. (c) shows the peak of the geoprofile in red, next to the now renamed Flower and Dean Street, a potential location in which the Ripper may have lived. Figure from Le Comber & Stevenson (2012).

1.6 From criminology to biology: applications of geographic profiling to biology

Although geographic profiling was originally designed to apply to crimes such as murder, rape and arson, its success in other areas, including burglary, counter-insurgency and piracy (see for example Kucera (2005) and Rossmo & Harries (2011)), its application to biological data was an obvious step.

Geographic profiling models present a new approach based on a simple spatial relationship. Two spatial patterns universally affect the dispersal and reproduction of any animal or plant. The first is distance decay; it takes effort to travel through space, and even those spores that hijack other animals or travel upon wind and air must contend with evolved responses or chaotic interactions that effectively limit their dispersal distance. The second principle is that of a buffer zone: few organisms can live around the same area as their offspring due to the competition for similar niche space that will inevitably result (sometimes only in an adult phase). The example of allelopathy in British trees is well documented (Singh et al. 2001). These simple relationships can be used to construct a mathematical model to predict the source of invasive populations

Given the similarities between criminal hunting behaviour and animal behaviour, it is perhaps not surprising that the first paper to apply geographic profiling in a biological context looked at animal foraging (Le Comber *et al.* 2006).

Geographic profiling and animal foraging

Geographic profiling was introduced to biology in a 2006 paper in the *Journal of*

Theoretical Biology (Le Comber e*t al.* 2006). In this study, the authors used data from radio-tracking studies of two species of bat, the common and soprano pipistrelles (*Pipistrellus pipistrellus* and *P. pygmaeus*) in north-east Scotland. A previous study had identified both roost sites and foraging sites, and the authors fitted Rossmo's criminal geographic targeting (CGT) model (Rossmo 2000) for each bat and showed that the fitted model parameters (B, f and g) could be used to locate roost sites, using foraging sites as input, analogous to crime sites. Interestingly, the fitted values differed between the two species, despite their close genetic relatedness. This probably reflects their different foraging strategies; *P. pygmaeus* forages preferentially in riparian habitats (ie along the edges of rivers and lakes) that support higher numbers of insect taxa (Gressit & Gressit 1962; Townes 1962), while *P. pipistrellus* is more generalist. This specialisation in *P. pygmaeus* is likely to mean that this species must forage over greater distances to locate sufficient prey items to satisfy its energetic demands. This was an intriguing result, suggesting that when anchor points such as nests, roosts or dens are known, fitted CGT model parameters could provide a concise way of describing complicated foraging patterns.

The bat study was followed by a second study of animal foraging, this time in bees, but using laboratory data rather than field data (Raine *et al.* 2009). Bees were allowed to enter a flight arena approximately 1m square, via a central hole, and allowed to forage on artificial flowers containing a sucrose solution. Again, the CGT algorithm successfully located this entrance. Fitting model parameters in the same way as in the bat study also showed that when the artificial flowers were presented at higher density, the size of the buffer zone decreased. This was of interest because, in criminology, little may be known about the target backcloth, since law enforcement agencies will have information on crimes committed, but not always on potential

crimes that were not.

Another interesting extension of this study involved using 'virtual' bees, in a similar experimental design to the real bees, using a variety of plausible foraging algorithms (including spiral searches, nearest-neighbour methods and a variety of others) (Raine *et al.* 2009). Just as the fitted model parameters could be used to differentiate between the foraging patterns of the two bat species, they could be used here to distinguish between different foraging rules (Raine *et al.* 2009). Crucially, for biologists, these could also be compared to the behaviour of the real bees, allowing the authors to rule out some of the suggested foraging algorithms as inconsistent with the patterns observed in the real bees (Raine *et al.* 2009).

At about the same time, Martin *et al.* (2009) used geographic profiling to study great white shark predation on seals off the coast of South Africa. Again, much of the interest of this paper derived from aspects tangential to the main purpose of geographic profiling; that is, identifying sources for point pattern data. In this case, the study identified a well-defined search base or anchor point 100 m seaward of the seals' primary island entry-exit point. This is not where the chances of intercepting seals are greatest, and the authors suggested that it represented a balance between prey detection, capture rates, and competition. In addition, the different geoprofiles observed for sharks of different ages showed that smaller sharks exhibited more dispersed search patterns and had lower success rates than larger sharks, suggesting either that hunting success improved with experience or that larger sharks excluded smaller sharks from the most profitable areas.

Geographic profiling and epidemiology

As noted above, animal foraging behaviour has much in common with criminal hunting behaviour. The extension of geographic profiling to epidemiological datasets, however, involves several important differences, notably the increased importance of multiple sources, and the possibility, for some diseases, of secondary sources. These issues are discussed below.

The application of geographic profiling to epidemiological data fills a surprising gap in epidemiology. As Buscema *et al.* (2009) pointed out, classical epidemiology tends to model the spread of infectious epidemic diseases, and few attempts have been made to identify the origin of the epidemic spread (excepting the classic case of John Snow and Cholera (Snow & Frost 1936)). This is surprising because, as Le Comber *et al.* (2011) noted, in many diseases infection sources can be highly clustered: for example, malaria parasite transmission is strongly dependent on the location of vector breeding sites, and most transmission only occurs within short distances of these sites; in Africa, these distances are typically between a few hundred meters and a kilometer, and rarely more than 2-3 km (Carter *et al.* 2000). Because of this clustering, untargeted control efforts are highly inefficient. Although source reduction of mosquito larval habitats can dramatically mitigate malaria transmission (Yohannes *et al.* 2005; Gu *et al.* 2006; Walker & Lynch 2007; Gu & Novak 2009), the transient nature and diversity of potential vector breeding sites makes the identification and control of breeding sites difficult (Carter *et al.* 2000). As a result, evidence-based targeting of interventions is more efficient, environmentally friendly and cost-effective than untargeted intervention. This, of course, is exactly the problem geographic profiling was designed to solve.

The first attempt to apply geographic profiling to epidemiogical data was by Buscema *et al.* (2009). This study examined Chikingunya fever, foot and mouth

disease and cholera, but concluded that geographic profiling was less efficient than the authors' preferred artificial intelligence method, the H-PST (Hidden-Pick and Squash Tracking) algorithm. However, this study mistakenly used the distance between the peak of the geoprofile and the correct source as a measure of model performance. As Rossmo (2000) was careful to point out, geographic profiling does not attempt to provide a point estimate for the anchor point (here, the infection source), as methods such as spatial mean, spatial median and centre of minimum distance seek to do; rather, it describes an optimal search strategy. Because of the complexity of jeopardy surfaces, the distance from the peak of the geoprofile to the anchor point is irrelevant; what matters is what percentage of points within the study area are higher on the geoprofile than the anchor point. In fact, when there are multiple sources of infection (eg the malaria cases examined by Le Comber *et al.* (2011)) this is an important advantage of geographic profiling over the H-PST algorithm, since methods that provide point estimates of sources will typically perform poorly when there is more than one source. When Le Comber *et al.* (2011) revisited one of the case studies in the Buscema paper (John Snow's data on the 1854 London cholera outbreak (Snow & Frost 1936)), geographic profiling performed extremely well.

Geographic profiling and invasive species biology

Another promising area for the application of geographic profiling to biological research concerns the spread of invasive species, an area with more in common with epidemiology (eg multiple and secondary sources) than with animal foraging. The issue is not trivial; invasive species are now viewed as the second most important driver of world biodiversity loss behind habitat destruction and have been identified as a significant component of global change (Vitousek *et al.* 1996; Wilcove *et al.* 1998). The cost of invasive species can run from millions to billions of dollars per occurrence (Baker 1986; Pimentel *et al.* 2001), and invasive species have been shown to affect native species through predation and competition, modify ecosystem functions, alter the abiotic environment and spread pathogens (Strayer *et al.* 2006; Ricciardi 2007). In addition, the problem is likely to get worse as climate change and anthropogenic influences lead to increased range shifts (Hulme 2007). For these reasons, prevention and control of invasive species has been identified as a priority for conservation organisations and government wildlife and agriculture ministries globally (Baker 1986; Hulme 2006). The application of GP to invasive species is discussed extensively in Chapter 2, in which a maximum likelihood approach to GP is developed and applied to data on 53 UK invasive species.

Geographic profiling is an exciting new approach that will work alongside and supplement existing methods and fits easily within the invasion stage model presented by Sakai et al. (2001). Niche modeling approaches may be incorporated easily into a geographic profile and it is clear that the relatively small amount of data needed to run the model is an advantage over more complex population dynamics based approaches. This new method could be used in the lag period and the expansive stages of the invasion to dramatically reduce the cost of controlling

invasive species by carefully targeting areas from which the problem species is expanding, whilst excluding those that are not.

Differences between biology and criminology

The first applications of geographic profiling to biology involved fairly straightforward mapping of the basic concepts from criminology: in these studies, animal foraging sites were used to identify animal roosts (or other home locations) in the same way that crime sites are used to identify probable areas of offender residence in criminology (Le Comber *et al*. 2006; Raine *et al.* 2009; Martin *et al.* 2009). However, later extensions, and most notably studies of invasive species biology and epidemiology, differ in a number of areas (Buscema *et al.* (2009); Le Comber *et al.* (2011); Stevenson *et al*. 2012) . In criminology, the application of geographic profiling will usually (or at least often) deal with the crimes of single individual with a single anchor point, often (hopefully!) over a short period of time. In contrast, biological data can involve multiple organisms (and hence multiple anchor points), secondary anchor points and extended time periods.

Multiple anchor points

In criminology, although jeopardy surfaces may have several peaks, relating perhaps to the criminal's home, work place or a relative's home (or, in the case of the Hillside Strangler, the two homes of the two cousins who committed the crimes together; Rossmo (2000)), it is usually assumed that the crimes are linked; that is, they are carried out by a single individual (some applications of geographic profiling

to terrorist activities may be an exception). In invasive species biology or epidemiology it is usually impossible, or at least impractical (perhaps requiring expensive genetic testing to identify particular strains of virus, or genotypes of individual plants or animals, for example), to link events to individual sources. For example, the malaria cases in Le Comber *et al.* (2011) were treated as a single group of 'crimes', although it is possible that six or more *Anopheles* spp. breeding sites were involved). In this case, data were simply pooled and the heights of each potential source on the geoprofile examined separately; Stevenson *et al.* (2012) took a similar approach with invasive species. At this point, no studies have explicitly examined the effect of multiple sources on geographic profiling model performance, although the data in Le Comber *et al.* (2011) and Stevenson *et al.* (2012), along with simulation data presented in Chapter 3, suggest that geographic profiling's performance relative to simple measures of spatial central tendency (spatial mean, spatial median, centre of minimum distance) will increase as the number of sources of increases. The issue of multiple sources is explicitly returned to in Chapter 3, where I present a model using Dirichlet process clustering that can locate multiple unknown sources from point pattern data.

Secondary anchor points

Murder victims do not go out and commit murders; victims of arson do not go out and burn down other buildings. Similarly, in the context of animal foraging, seals predated upon by great white sharks do not then predate upon other seals. However, the sites of new biological invasions can go on to act as sources for further waves of invasion; similarly, in many disease systems, infected individuals will go on to infect

other individuals. These secondary sources/anchor points may dramatically alter the spatial patterns observed.

Extended time periods

In criminal investigations, the persistence of a series of linked crimes over a number of years obviously represents a failure of law enforcement; cases such as Jeffrey Dahmer (1978 to 1991) or the Yorkshire Ripper (1975 to 1980) (Rossmo 2000) are, hopefully, an exception. In biology, this need not be the case, and longer-term datasets may in fact be highly desirable. Ecological data sets in particular can span decades or even centuries (Stevenson *et al.* 2012), and can involve multiple 'outbreaks', while criminal cases typically span shorter periods of time. In this sense, biological data may offer a distinct advantage over criminological data. Invasions and disease outbreaks have long histories and repeated outbreaks, so assuming that repeated invasions follow similar histories, previous outbreaks (perhaps with known sources) can be used to validate the geographic profiling model (Stevenson *et al.* 2012). Future spread can then be predicted using parameters established from the organism's own invasion history (Stevenson *et al.* 2012). The application of GP to invasion biology is discussed in depth in Chapter two.

Conclusions

The application of GP to biological data is new but shows great promise. GP has been shown to work in a few cases such as animal foraging that are very similar in scope to the original use of GP in criminology (Rossmo & Harries 2011). GP has

also been applied to the problem of locating sources of malaria infection. This is a significant development in the field and takes GP far outside of its normal area of applicability (Le Comber *et al.* 2011). GP continues to demonstrate its effectiveness and robustness to a range of problems (Rossmo & Harries 2011). There are key differences and areas that need exploring for GP to be effective in biology but these are small issues compared to the massive potential benefit of bringing a new approach to our field.

1.7 Conclusions and discussion of research direction

The GP model, originally developed to solve the problem of locating serial offenders, has grown far beyond its original scope. It is applied to a wide range of problems in criminology and in military science. It has also been used in animal foraging and epidemiology and can work together with other models within an ecological framework. The previous section should serve to illustrate the strengths of the model, which are: (i) practicality and ease of use; (ii) its diversity of application; (iii) its intuitive nature, and (iv) its proven success in a range of fields. Weaknesses of the model are: (i) to date, limited application in ecological/biological settings; (ii) little formal testing with simulated data to date, and (iii) its lack of underlying mathematical theory.

In the following chapters I will address these weaknesses and expand on the existing strengths. Specifically, in Chapter 2 I will begin to address the mathematical problems by discussing model fitting, and then test this new approach with real and simulated data; the bulk of the work in this chapter is published as Stevenson *et al.* 2012. In Chapter 3 I will look in depth into the mathematical and theoretical issues

in the GP field and derive a new Bayesian approach. In Chapter 4 I will run further simulations comparing GP to other models. In Chapter 5 I derive functional dispersal profiles from ecological data, allowing GP to be applied easily to biological problems. Finally, in Chapter 6, I use the fully formed GP methods in three widely different examples from three separate fields. This thesis will take GP from an obscure method used extensively in an alien field to an up-to-date approach applicable to a wide range of biological problems.

Chapter 2: Fitting and testing the Rossmo function

[Chapter 2 closely follows Stevenson et al. 2012]

2.1 Abstract

This chapter has two main parts. In the first, I use computer simulations to compare GP to other simple measures of spatial central tendency (centre of minimum distance, spatial mean, spatial median), as well as to a more sophisticated single parameter kernel density model. GP performs significantly better than any of these other approaches. In the second part of the study, I analyse historical data from the Biological Records Centre (BRC) for 53 invasive species in Great Britain, ranging from marine invertebrates to woody trees, and from a wide variety of habitats (including littoral habitats, woodland and man-made habitats). For 52 of these 53 data sets, GP outperforms spatial mean, spatial median and centre of minimum distance as a search strategy, particularly as the number of sources (or potential sources) increases. I analyse one of these data sets, for *Heracleum mantegazzianum*, in more detail, and show that GP also outperforms the kernel density model. Finally, I compare fitted parameter values between different species, groups and habitat types, with a view to identifying general values that might be used for novel invasions where data are lacking. I suggest that geographic profiling could potentially form a useful component of integrated control strategies relating to a wide variety of invasive species

2.2 Introduction

As noted in Chapter 1, invasive species are a major cause of world biodiversity loss, with costs reaching billions of dollars in some cases (Vitousek *et al.* 1996; Wilcove *et al.* 1998; Mooney & Drake 1986; Pimentel *et al.* 2000). Here I ask whether GP can be used to identify source populations of invasive species as a prelude to targeted control. Crucially, if geographic profiling (GP) is to provide an effective method for locating source populations of invasive species, it first of all needs to be demonstrated that it can do better than other methods of prioritising searches. In addition, since it is likely that many or even most invasions will concern species where data are, at least initially, lacking, it would clearly be helpful if general model values could be used for different groups of taxa (for example, what values of the buffer zone radius are most appropriate for woody trees, or marine invertebrates?). In this study, I address both of these issues. Specifically, I ask (i) Can geographic profiling can be used to locate sources of invasive species? (ii) If so, is GP more efficient than other simple approaches such as spatial mean, spatial median or centre of minimum distance, or a more complex single parameter kernel density model? (iii) Is it possible to use general parameter values for different species or groups or for species occupying different types of habitats in cases where data may be lacking? (iv) Do model variables alter over time, specifically in the earlier or later stages of invasions?

2.3 Methods

General approach

My general approach was to fit the GP model parameters using species locations at time t in such a way that they optimally predict species locations at time t–1, and

then validate the model by using the same fitted parameters to test whether the locations at time t–1 predict the locations at time t–2. In this way, model fitting and model testing are independent. In this case, the time steps chosen were decades (eg 1960-69, 1970-79). For each of the species analysed, three decades were chosen (times t, t–1 and t–2). The decades selected were not the same for each species, since decades where there were large data sets were preferentially chosen; early datasets, where data might be expected to be unreliable and susceptible to differences in sampling effort, were also avoided.

Geographic profiling model

A full description of the model can be found in Section 1.5. I used the full Rossmo approach (Rossmo 2000) with Euclidian distance introduced by Le Comber *et al.* (2006). The model was implemented using the R statistical package (R development core team 2012).

Model fitting

To validate the model, I took advantage of the temporal resolution of data at the Biological Records Centre (see 'Spatial data', below) to split species data into decades (eg 1960-1969, 1970-1979, 1980-1989). I then fitted the model by selecting the value of B that best predicted the locations of the species in question in (for example) 1970-79, using as input the locations from 1980-1989. This was done using a maximum likelihood approach with quasi-newton optimisation (Nocedal & Wright 1999) to fit B, and leaving f and g fixed at 1.2. These are the values typically

used in criminology, and previous studies have shown that the model is much more sensitive to changes in *B* than to changes in *f* and *g* (Le Comber *et al.* 2006; Raine *et al.* 2009). This also helps to avoid the problem of overfitting as discussed by O'Leary (2009) in the context of geographic profiling, which can result in a model that essentially reduces to a series of point estimates that perfectly predict the data used to fit the model, but that have little or no predictive power when applied to other datasets. A further advantage, in the context of this study, is that this allows the model to be constrained to a single parameter allowing direct comparison with a single parameter kernel density model.

The model's degree of fit can be calculated using the hit score percentage (HS%), the proportion of the area covering the crimes (in this case, the locations of invasive species) in which the offender's base (or the source of the invasive species) is located; in criminology, this is usually the area bounding the crimes, plus a 'guard rail' of 10% surrounding this. The HS% is calculated by dividing the ranked score (p_{ij}) by the total search area and multiplying by 100. The smaller the HS%, the more accurate the geoprofile; a hit score of 50% is what would be expected from a nonprioritised (i.e., random or uniform) search (Rossmo 2000). In our analysis, unlike criminology, there are multiple sources, so I calculated the mean hit score percentage across all locations. This 'Learning hit score' is reported in Table 2.1.

Model testing

To test the model's performance, I used the fitted parameters to test the model's predictions on an earlier time step; in the example above, I would feed the locations in 1970-79 back into the model, and calculate the hit score percentages of the

locations in the previous time step (1960-69) (the test hit score in Table 2.1), thus ensuring that the learning and test hit score percentages were independent.

Other measures of spatial central tendency

GP was compared to other simple measures of spatial tendency commonly used in both biology and criminology. The two most basic approaches are to calculate the spatial mean (or centroid) and the spatial median. The spatial mean is the point where the coordinates are the mean of the x and y coordinates. The equation is shown below:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

(Equation 2.1)

where x_i and y_i are the co-ordinates of the invasive species presence and *n* is the total number of locations where the species is present. The spatial median is the median point of the x and y co-ordinates. Using the same notation as above it is calculated as follows:

$$\bar{x} = \frac{1}{2}\left(x_{\lceil \frac{n}{2} \rceil} + x_{\lfloor \frac{n}{2} \rfloor + 1}\right),\ \bar{y} = \frac{1}{2}\left(y_{\lceil \frac{n}{2} \rceil} + y_{\lfloor \frac{n}{2} \rfloor + 1}\right)$$

(Equation 2.2)

A slightly more complex approach is to use the centre of minimum distance (CMD). The CMD is the location at which the sum of the distances to all other points is minimised and is described by the following equation:

$$W(\bar{x}, \bar{y}) = \sum_{i=1}^{n} \text{dist}((x_i, y_i), (\bar{x}, \bar{y}))$$

(Equation 2.3)

W is the distance from each location of an invasive species (x_i, y_i) to the chosen point ($\bar{x}$, $\bar{y}$). W is then minimised using an iterative algorithm. The function of W is a convex hull and as such has a unique minimum. I used the algorithm developed by Weiszfeld (1936):

$$(x^{(k+1)}, y^{(k+1)}) = \left(\frac{\sum_{i=1}^{n} x_i/\text{dist}((x_i, y_i), (x^{(k)}, y^{(k)}))}{\sum_{i=1}^{n} 1/\text{dist}((x_i, y_i), (x^{(k)}, y^{(k)}))}, \frac{\sum_{i=1}^{n} y_i/\text{dist}((x_i, y_i), (x^{(k)}, y^{(k)}))}{\sum_{i=1}^{n} 1/\text{dist}((x_i, y_i), (x^{(k)}, y^{(k)}))}\right)$$

(Equation 2.4)

where the initial values of x and y may be taken as any reasonable value (the centroid for example). The method has been shown to always converge on the CMD, yet the length of time taken for this to occur is an unknown function dependent on the size and distribution of the data (Kuhn & Kuenne 1962). I ran the method for k = 100 to ensure its convergence.

These methods were then compared to GP by searching outwards from the central point in concentric bands. Using this search strategy, a figure comparable to the hit score could be calculated based on a directed search originating out from these measures of spatial central tendency.

Kernel density method

In addition to the simple methods described above, I also compared GP to a kernel density method of the type that underlies gravity models. In this study, I adapted the kernel density method for estimating home range size described by Worton (1989) which underlies the backcasting application in MacIsaac *et al.* (2004). This kernel density approach uses a fixed kernel based on the summation of unimodal bivariate normal probability density functions. Whilst similar in some ways to the geographic profiling model, this approach uses a single parameter normal distribution in place of the geographic profiling function. The single parameter bivariate normal kernel is given by the following equation (note that the notation used in this equation, but not the form, have been modified to make the comparison with the GP model more explicit):

$$p_{ij} = \frac{1}{nh^2} \sum_{i=1}^{n} \frac{1}{2\pi} . exp\left(-\frac{(x_i - x_n)'(y_i - y_n)}{2h^2} \right)$$

(Equation 2.5)

where p_{ij} is the probability of each point being a source. (x_i, y_i) and (x_n, y_n) are the same as in the geographic profiling model. *h* is the only free parameter in the model and can be used to smooth out or concentrate the surface; it is referred to as the smoothing parameter. This parameter may be fitted in much the same way as the geographic profiling model's parameter *B* as discussed above.

Simulations

Simulated data were also created to test and compare the GP model alongside real species invasions. The simulations were created using the statistical modelling environment R (R development core team, 2008), and consisted of a 100 by 100 grid in which sources were uniformly distributed in the central 36*36 region (the constraint was necessary to avoid edge effects). From each source a normal distribution of spread points was simulated with a standard deviation of 5. The simulations tested every possible combination of 1, 2 and 5 sources and 10, 20 and 30 spread points. Each variation was replicated 1000 times and the hit score for each model calculated to test how well they identified the source locations. Where there were multiple sources, the mean hit score of all of the individual sources was used.

Two sets of simulations were conducted. In the first, GP was tested against the spatial mean, spatial median and centre of minimum distance. In this set of

simulations (as in most real-world examples), the dispersal parameters are unknown, and GP uses half the mean nearest-neighbour distance between points to estimate B. In the second set of simulations, GP was compared to the kernel density method. The kernel density method uses as an input h, which corresponds to the standard deviation of the normal distribution used to simulate dispersal, and in this set of simulations both h and B were set to the true standard deviation.

Spatial data

53 British invasive species were chosen that had extensive invasion histories as recorded in the R662 audit of non-native species of England, produced by Natural England (Hill *et al.* 2005). The R662 audit was a desk-based survey that collated all of the standard British resources and checked these against up-to-date studies. Data on the locations and date of presence (taken as the centroid within $10km^2$ grid) were obtained from the Biological Records Centre (BRC) database downloaded from NBN gateway (http://www.nbn.org.uk/). The BRC database contains datasets from hundreds of contributors across the UK, ranging from government bodies, NGOs and scientific surveys. There is significant variation in the quality of the data collected, but all species chosen had multiple records collected from different organisations. A full list of the species chosen, years analysed and results obtained for each species can be found in Table 2.1.

Taxonomic differences

Different species have very different reproductive habits and dispersal traits, and this

is likely to be reflected in differences in the radius of the buffer zone (*B*) in our model. To test for differences in *B* between species belonging to different taxa, I examined the distribution of *B* values in major and minor taxonomic functional groups as defined by English Nature's *Audit of non-native species* (Hill et al. 2005).

Habitat differences

Different species have preferred habitat types and the abundance of suitable habitats, and their distribution may affect the rates at which species are able to spread. In its current formulation, GP makes no attempt to include habitat information, but it is still possible to check whether there is an association between the values of *B* produced by species living in different habitats. I examined *The European Nature Information System* (EUNIS) habitat types for all 53 of the species analysed to test for patterns in *B* values associated with habitat differences. The EUNIS habitat classification is a pan-European system, which was developed between 1996 and 2001 by the European Environment Agency (EEA) in collaboration with experts from throughout Europe, and covers all types of natural and artificial habitats, both aquatic and terrestrial (Davies *et al.* 2004).

Temporal differences

Invasions can span decades or even centuries. In this area the biologist has an advantage over the criminologist, since data can be divided into different temporal units, allowing examination of changes in model parameters as invasions progress. Invasions can change over time, and the notion of an invasion delay, followed by an

explosion of expansion, is well established (Crooks & Soulé 1999). I selected three species with extensive invasion histories and repeated the analysis over multiple decades. In this way I could determine if the fit of the model alters over the course of an invasion history and could also compare species to determine if they had consistently different B values across time.

2.4 Results

Simulations

The mean hit scores for each combination of conditions used in the simulation are presented in Table 2.2 and Figure 2.1,

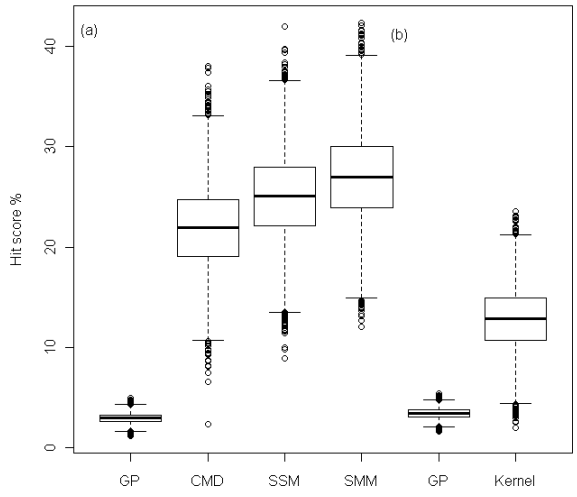


Figure 2.1 Boxplot of simulation results, with GP compared to (a) centre of minimum distance (CMD), spatial mean (SMM) and spatial median (SSM), and (b) a kernel density model. Mean hit score % across all 9000 runs of each method (1000 replicates x three levels of source points x three levels of spread points) are shown on the y axis. GP performs well across all tests and never took more than 7% of the search area to find all of the sources. All other methods show much greater range

in success. The kernel density method performs well when number of sources equals 1 but then rapidly begins to decline in efficiency as the number of sources increase.

The results were analysed using a full factorial ANOVA with model type, number of sources and number of spread points as factors, and all interactions included. The first test compared GP to spatial mean, spatial median and centre of minimum distance. The results showed significant differences between model type, but not between other factors or interactions (model type: $F_{3,35984} = 75386.2$, $p < 2e\text{-}16$; number of sources: $F_{1, 35984} = 3.6$, $p = 0.056$; number of spread points: $F_{1, 35984} = 1.4$, $p = 0.23$; model type x number of sources: $F_{3, 35984} = 0.5$, $p = 0.66$; model type x number of spread points: $F_{3, 35984} = 0.1$, $p = 0.94$; number of sources x number of spread points: $F_{1, 35984} = 0.0$, $p = 0.89$; model type x number of sources x number of spread points: $F_{3, 35984} = 2.1$, $p = 0.1$). Tukey post-hoc tests showed that GP performed significantly better than all other methods in each case (Table 2.2a). The second test compared GP to the kernel density method. Again, the results showed significant differences between the models (model type: $F_{1,17992} = 81216.9$, $p < 2e\text{-}16$; number of sources: $F_{1,17992} = 0.1$, $p = 0.75$; number of spread points: $F_{1,17992} = 2.3$, $p = 0.13$; model type x number of sources: $F_{1,17992} = 0.2$, $p = 0.67$; model type x number of spread points: $F_{1,17992} = 2.7$, $p = 0.19$; number of sources x number of spread points: $F_{1,17992} = 0.1$, $p = 0.32$; model type x number of sources x number of spread points: $F_{1,17992} = 0.4$, $p = 0.55$). Tukey post-hoc tests showed that GP performed significantly better than all other methods except when the number of sources was 1, when the kernel density model performed as well as GP (Table 2.2b).

Model performance

Across each of the 53 species examined, GP's mean hit score percentage was lower than the 50% that characterises a random search (mean ± sd: 18% ± 18.4%; log transformation: t = 37.338, df = 52, p < 2.2e-16). GP also outperformed the spatial mean, spatial median and centre of minimum distance in 52 out of 53 datasets. GP had a mean hit score percentage of 18.4% across all datasets (SD = 6%) compared to 58.1% (SD = 14%) for the spatial mean, 48.7% (SD = 9%) for the spatial median and 43.4% (SD = 9%) for the CMD. ANOVA reveals significant differences between these ($F_{3,50}$ = 100.01, p < 2.2e-16) (Table 2.2).

I also selected one dataset, *Heracleum mantegazzianum*, for which there are good data, for more detailed analysis, examining the distribution of hit scores for individual sites, as well as considering the average across all sites (Figure 2.2). For 35 of 51 1941-1950 sources the hit score percentage was below 10%. The mean hit score percentage for all the sources was 13% and never exceeded 24% of the searchable area. I also ran a kernel density model on the *Heracleum mantegazzianum* data set, to compare its performance to that of GP. For 18 of 1941-1950 sources the hit score percentage was below 10%. The mean hit score was 31% and reached a maximum of 65% of the search area. This result was significantly worse than that of GP (Wilcoxon rank sum test: W= 382, n = 51, p = 8.052e-10).

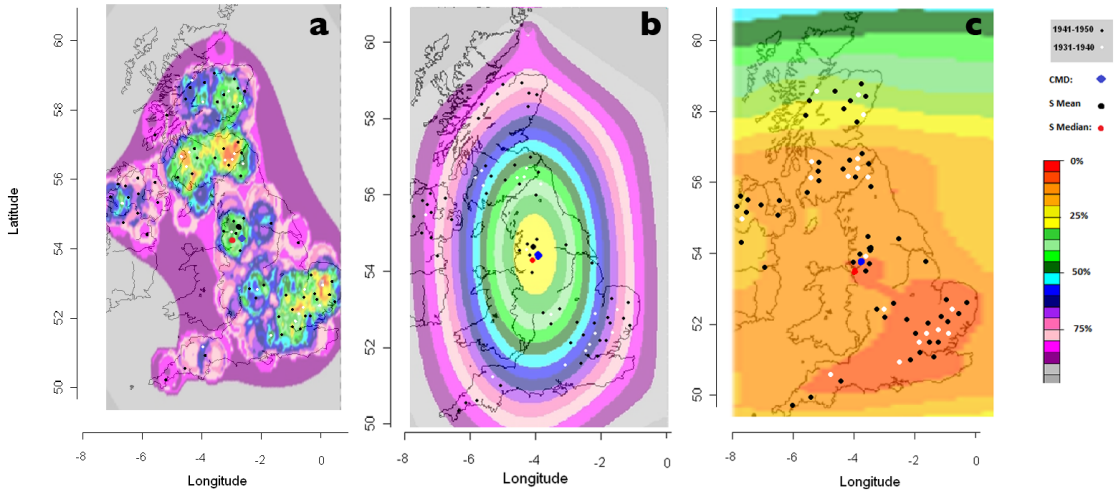


Figure 2.2 Search strategies for *Heracleum mantegazzianum*, based on (a) geographic profiling; (b) centre of minimum distance, and (c) kernel density model. Black circles show the locations in 1941-50, and white circles the locations in 1931-1940. Contours are shown in bands of 5%. Geoprofiling successfully locates 35 of 51 source populations after searching just 10% of the target area, with a mean hit score for all sources of 13%. Other approaches do not perform as well. The centre of minimum distance (CMD), spatial mean (S mean) and spatial median (S median) are also shown in blue, black and red respectively.

Taxonomic differences

There were no significant differences in fitted *B* values across the English Nature major groups animals, marine, plants (mean ± sd: animals: 0.43 ± 0.08; marine 0.56 ± 0.48; plants 0.42 ± 0.18; log transformation: $F_{4,49} = 0.60$, $p = 0.44$), although I noted that the variance for the category 'animals' was lower than in the other two categories. However, when I looked within the category plants to compare categories 76 (Woody stemmed conifers) and 77 (Deciduous flowering plants), I did find significantly different values of *B* (median values: category 76: 0.17; category 77:

0.16; Wilcoxon rank sum test: W = 238, n = 11, 30, p = 0.005) (Figure 2.3).

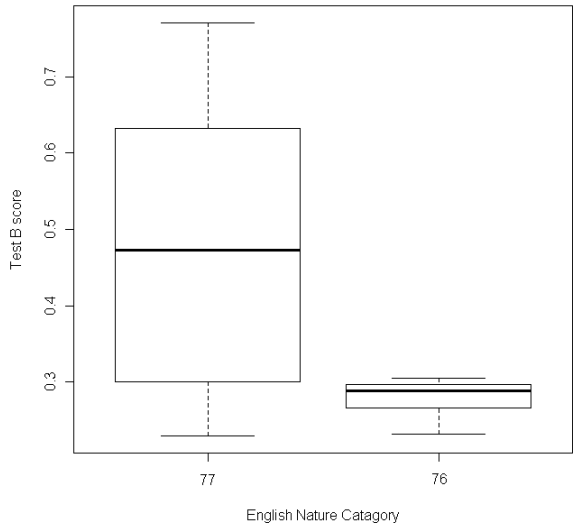


Figure 2.3 Boxplot showing fitted values of *B* for English Nature category listings 76 (Conifers and gingko) and 77 (Flowering plants).

Habitat differences

When I fitted the model to different EUNIS habitat listings, I found significant differences in fitted values of B (log transformation: ANOVA $F_{4,46} = 3.26$, $p = 0.02$), although this was driven largely by differences between categories G (Woodland or forest) and J (Industrial, constructed or artificial habitats) (Figure 2.4).

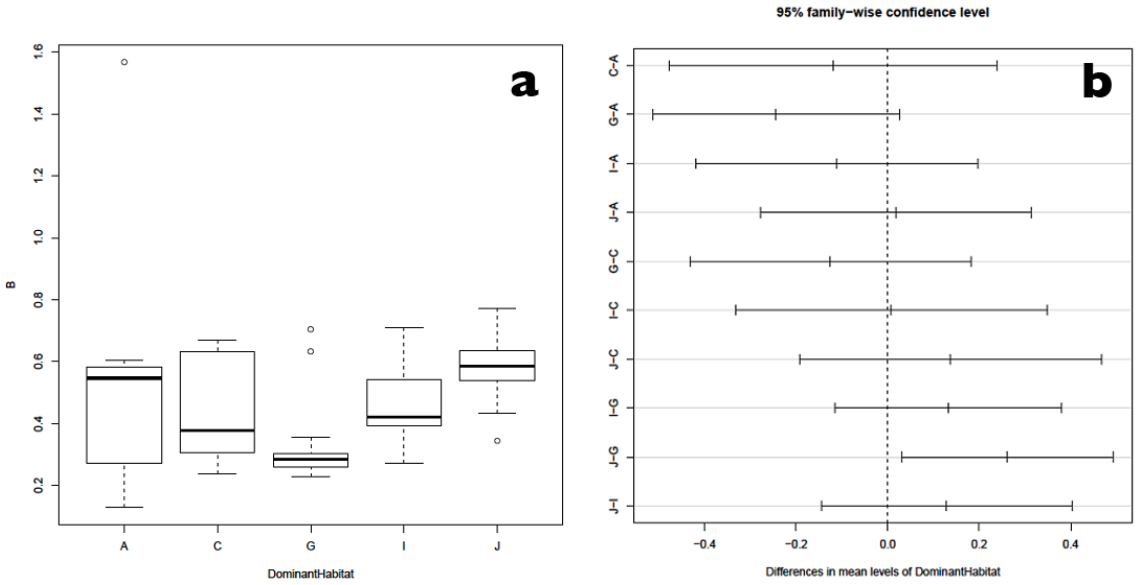


Figure 2.4 Boxplot showing (a) fitted values of B for EUNIS habitat categories A (Marine habitats), C (Inland surface waters), G (Woodland, forest and other wooded land), I (Regularly or recently cultivated agricultural, horticultural and domestic habitats) and J (Constructed, industrial and other artificial habitats); (b) Tukey 95% post-hoc comparisons for each of these categories, showing that only categories J and G are significantly different.

Temporal differences

For the three species for which I analysed multiple time periods, fitted values of *B* did not differ within species across different time steps (linear regressions of slopes were not significantly different from 0 in all cases). However, the method detected significant differences in fitted values of *B* between individual species ($F_{2,20}$ = 8.5248, p = 0.0005474) (Figure 2.5).

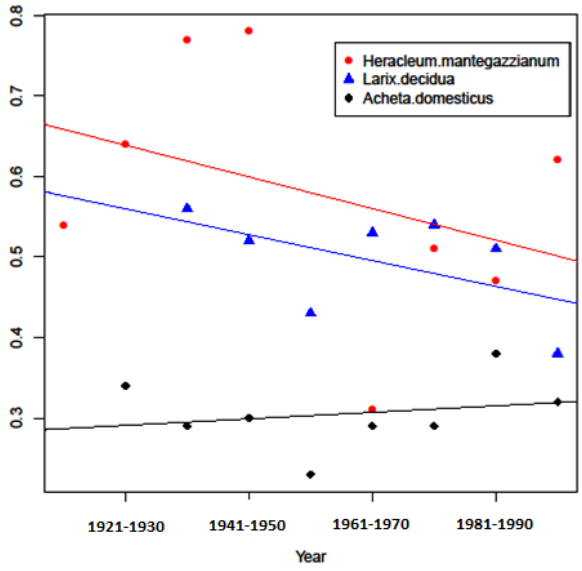


Figure 2.5 Time series plots for *Heracleum mantegazzianum*, *Larix decidua* and *Acheta domesticus*. In none of these three cases did fitted values of *B* differ within species across different decades (linear regressions were non-significant in all cases). However, the method detected significant differences in fitted values of *B* between individual species ($F_{2,20}$ = 8.5248, p= 0.0005474).

2.5 Discussion

Our study shows that geographic profiling can correctly predict the sources of invasive species, using as input their current locations. Crucially, it can also do so in the early stages of invasions with data that is possible to acquire quickly, and when control efforts are most likely to be effective. Geographic profiling outperformed other widely used spatial statistics such as the centre of minimum distance, spatial mean and spatial median in locating invasion sources. In addition, a simple kernel density approach was found to be less effective than GP, with GP's advantage increasing as the number of sources of invasion increased. This study also suggests that it may be possible to use general, taxon- or habitat-specific values for the model parameters in cases where data on individual species are lacking, since different

fitted values of B were obtained in flowering plants and conifers.

There are marked similarities between the problems faced by police investigators and invasion managers. Both can face situations with large volumes of data and difficulty in determining which data are informative and which uninformative. In both cases, resources and time are limited (Williamson & Fitter 1996). Methods for optimising searches for source locations of invasive organisms could therefore provide a valuable tool in the rapid assessment and control in the critical early stage of invasion history (Leung *et al.* 2002), just as they have done in criminology (Rossmo 2000).

Many invasive organisms show distinctive time lag between initial invasion, establishment and subsequent expansion (Crooks & Soulé 1999). Often repeated invasions will occur by the same pathway and only after a period of time will one of these invasions begin to expand beyond the initial source of invasion (Drake & Lodge 2004). GP operates at this critical early stage of the invasion process, locating the source locations of rapidly expanding populations allowing effective allocation of limited resources available for control measures (Keller *et al.* 2008; Puth & Post 2005).

Following the work of numerous scientific workers such as weed scientists, resource managers, conservation biologists, restoration biologists, field ecologists and economists, a clear understanding of the stages that make up invasions and the relevant modelling and available management options to prevent and manage invasive species (Sakai *et al.* 2001; Williamson & Fitter 1996b), Sakai *et al.* (2001) present an example invasion framework (modified from Lodge (1993)) that highlights the general steps in the invasion process and their relationship to

management steps that can be taken. The movement, establishment and subsequent spread of invasive species are best characterised by a series of discrete steps, each of which poses different problems to both manager and modeller (Vitousek 1997). Some of these stages are more relevant to prevention; others are more relevant for issues of control and restoration. There has been increasing understanding that feedback may occur between many of these steps (Sakai *et al.* 2001). Within each of the phases, different types of predictive and analytical models can be used to gather information and make assessments of the risks presented by invasive species. The results reported in this study suggest that geographic profiling could form part of an integrated strategy and allow more precise targeting of source populations and thus more effective allocation of resources. This will be most useful in the establishment and lag phase, but may still be useful in populations that have started to expand rapidly, since the model may be used to differentiate between existing populations that are acting as sources and those that are not, again improving the efficiency of intervention.

GP models present a new approach based on two simple spatial concepts, distance decay and the buffer zone. Distance decay has obvious relevance to the spread of invasive species, since movement involves costs, either in terms of energy expenditure or exposure to risk of predation, both of which limit dispersal distance. The buffer zone applies even without active avoidance of nearby sites since, if suitable habitats are randomly distributed, the number of sites increases geometrically with distance from the source. In fact, there may be cases where buffer zones arise from aspects of a species' biology. In some species, offspring avoid occupying the same area as their parents due to the competition for similar niche space that will inevitable result (sometimes only in the adult phase); for example,

allelopathy in British trees (Singh *et al.* 2001). There is also evidence for buffer zones in bee species (Dramstad 1996; Saville *et al.* 1997).

GP models operate in a similar fashion to gravity and kernel density methods, but have the distinct advantage of performing well with multiple sources of invasion, as well as being possible to run quickly and easily on a small number of data points; crucially, GP models run using only presence/absence data, without the need to estimate parameters for mechanisms such as outflows and inflows between sources and destinations, risk ranks etc. Rossmo (2000) has shown using Monte Carlo simulations that GP is capable of producing reliable profiles with as few as five data points.

The results presented here demonstrate the feasibility of the method, but there is also significant development potential. In two key areas there is immediate improvement that can be made in the model. First, GP can be easily placed within existing modelling frameworks in invasion biology. GP models can work alongside population growth models and be informed by trait-based risk analysis (Leung *et al.* 2002). One exciting area for future research would be to integrate GP with niche-based modeling (Peterson 2003; Thuiller *et al.* 2005). GP models are based on data that describe where species are now and how they have historically spread, a spatial relationship that does not include any information on preferable habitat types. Because niche-based models do not include any spatial data but are based upon habitat preferences (Leung *et al.* 2004; Thuiller *et al.* 2005), these two modeling types could be combined to produce a risk map that incorporates where species are, how they spread and how likely they are to settle in certain areas.

Second, the mathematics of GP is under continual development and is discussed by

O'Leary (2009). There is some controversy about which is the best model to use and how parameters are to be fitted (Canter *et al.* 2000; Levine 2009). The criticisms levelled by O'Leary (2009) over the lack of mathematical robustness are strong. He proposes a movement into a Bayesian framework and I agree with this, at least from the point of view of model testing and mathematical robustness. Future work will compare different modelling approaches in terms of model fit, ease of use and practical application as well as continuing to assess its use in biological systems.

2.6 Conclusion

I suggest that invasive species managers and conservation biologists should strongly consider the use of GP, especially when it is likely that there are multiple source populations of the species in question. Although GP and kernel models perform equally well when there is only a single source, our computer simulations show that, as the number of sources increases, simple kernel density models rapidly become less effective (for example, with five sources and 30 data points (a more than reasonable biological case), kernel models searched 13% of the area before finding all the sources, while GP searched on average 2.6% of the area before finding all sources). Managers are cautioned that using kernel models in the case of multiple sources could lead to searching/targeting significantly more of the target area than is necessary, involving a corresponding increase in effort. In real-world examples, where resources are likely to be limiting, geographic profiling offers an increase in search efficiency over other methods – such as spatial mean, spatial median, centre of minimum distance and simple kernel density models – that is likely to lead to improved targeting of interventions, and more efficient use of scarce resources.

Table 2.1 Animal and plant species used in this analysis, with English Nature category listings and EUNIS habitat data. EUNIS habitats are as follows: A1 Littoral rock and other hard substrate; C3 Littoral zone of inland surface waterways; G1 Broadleaved deciduous woodland; G4 Mixed deciduous and coniferous woodland; I1 Arable land and market gardens; I1 Cultivated areas of gardens and parks; J2 Low density buildings; J4 Transport networks and other constructed hard-surfaced areas; NA not applicable. 'Time step' shows the data used to fit the model, prior to testing. B is the fitted value of the buffer zone radius. The learning hit score is shown, along with the test and, for comparison, hit scores for the centre of minimum distance (CMD), spatial median (SSM) and spatial mean (SMM). For each species analysed, the most effective of these four search strategies is shown as a shaded cell; in 52 of 53 (98%) cases, GP out-performed all three other strategies. Data from English Nature audit of non-native species (Hill et al. 2005).

Species	Common name	English Nature listing category	Major category	Minor category	Dominant habitat	Time step	B	GP learning hit score	GP test hit score	CMD test hit score	SSM test hit score	SMM test hit score
Acheta domesticus	House cricket	55	Animals	Insects	NA	1971-1980; 1981-1990	0.38	7 00E-69	**0.25**	0.525	0.581	0.579
Aglossa pinguinalis	Large tabby	53	Animals	Insects	I2	1981-1990; 1991-2000	0.42	5 29E-06	0.23	0.471	0.654	0.557
Chrysalina americana	Rosemary beetle	52	Animals	Insects	NA	1991-2000; 2001-2010	0.35	3 72E-24	0.14	0.423	0.830	0.431
Dicranopalpus ramosus	Harvestman	44	Animals	Invertebrate non-insect	I2	1981-1990; 1991-2000	0.44	3 14E-94	0.12	0.335	0.469	0.373

Pholcus phalangioides	Daddy-long-legs spider or Cellar spider	44	Animals	Invertebrate non-insect	I2	1991-2000; 2001-2010	0.58	6 60E-08	0.22	0.485	0.563	0.514
Tegenaria agrestis	Hobo spider	44	Animals	Invertebrate non-insect	I2	1991-2000; 2001-2010	0.39	1 35E-71	0.26	0.519	0.600	0.508
Corophium sextonae	-	14	Marine	Marine & estuarine invertebrates	A1	1981-1990; 1991-2000	0.36	1 77E-27	0.24	0.466	0.588	0.578
Crassostrea gigas	Pacific oyster or Japanese oyster	13	Marine	Marine & estuarine invertebrates	A1	1981-1990; 1991-2000	0.6	4 88E-07	0.16	0.394	0.843	0.479
Crepidula fornicata	Common slipper shell	13	Marine	Marine & estuarine invertebrates	A1	1971-1980; 1981-1990	0.19	3 96E-63	0.18	0.467	0.480	0.470
Elminius modestus	Barnacle	14	Marine	Marine & estuarine invertebrates	A1	1971-1980; 1981-1990	0.55	7 20E-41	0.14	0.373	0.448	0.391
Petricola pholadiformis	False angel wing American Piddock	13	Marine	Marine & estuarine invertebrates	A1	1981-1990; 1991-2000	0.56	1 70E-14	0.21	0.446	0.562	0.523
Styela clava	Stalked or leathery sea squirt	18	Marine	Marine & estuarine invertebrates	A1	1981-1990; 1991-2000	0.13	1 32E-22	0.21	0.449	0.537	0.481
Undaria pinnatifida	Japanese kelp	33	Marine	Marine & estuarine plants	A1	1981-1990; 1991-2000	1.57	7 00E-69	0.18	0.454	0.545	0.522

Acer platanoides	Norway maple	77	Plants	Vascular plants	G1	1951-1960; 1961-1970	0.27	1 35E-48	0.25	0.455	0.573	0.568
Aesculus hippocastanum	Horse chestnut	77	Plants	Vascular plants	G1	1931-1940; 1941-1950	0.71	4 98E-43	0.25	0.461	0.557	0.551
Anisantha sterilis	Barren brome	77	Plants	Vascular Plants	I1	1971-1980; 1981-1990	0.27	5 34E-29	0.27	0.529	0.668	0.592
Campanula poscharskyana	Trailing bellflower	77	Plants	Vascular plants	J2	1971-1980; 1981-1990	0.59	1 14E-20	0.21	0.414	0.930	0.525
Ceratochloa marginata	Western brome	77	Plants	Vascular plants	C3	1971-1980; 1981-1990	0.63	1 36E-20	0.32	0.569	0.716	0.613
Cryptomeria japonica	Japanese red cedar	76	Plants	Vascular plants	G4	1951-1960; 1961-1970	0.29	4 71E-21	0.11	0.743	0.920	0.853
Erinus alpinus	Fairy foxglove	77	Plants	Vascular plants	J2	1971-1980; 1981-1990	0.64	8 68E-13	0.14	0.419	0.483	0.398
Fallopia japonica	Japanese knotweed	77	Plants	Vascular plants	C3	1931-1940; 1941-1950	0.38	5 29E-09	0.19	0.418	0.576	0.491
Fuchsia magellanica	Fuchsia	77	Plants	Vascular Plants	I1	1931-1940; 1941-1950	0.42	9 98E-45	0.14	0.354	0.755	0.411

Fumaria reuteri	Martin's ramping-fumitory	77	Plants	Vascular Plants	I1	1981-1990; 1991-2000	0.54	1 34E-58	0.12	0.330	0.465	0.458
Galanthus nivalis		77	Plants	Vascular plants	G1	1931-1940; 1941-1950	0.26	4 14E-17	0.19	0.441	0.549	0.483
Geranium dissectum	Cut-leaved crane's-bill	77	Plants	Vascular Plants	I1	1931-1940; 1941-1950	0.71	7 68E-25	0.19	0.401	0.510	0.503
Heracleum mantegazzianum	Giant Hogweed	77	Plants	Vascular plants	J4	1931-1940; 1941-1950	0.77	0	0.13	0.401	0.427	0.411
Juglans regia	Common walnut, Persian walnut or English walnut	77	Plants	Vascular plants	G1	1971-1980; 1981-1990	0.23	6 65E-71	0.13	0.401	0.520	0.423
Kickxia spuria	Roundleaf cancerwort	77	Plants	Vascular plants	I1	1951-1960; 1961-1970	0.3	7 83E-42	0.11	0.385	0.484	0.377
Lactuca serriola	Prickly lettuce	77	Plants	Vascular plants	J4	1931-1940; 1941-1950	0.35	3 45E-11	0.16	0.123	0.211	0.234
Larix decidua	European larch	76	Plants	Vascular plants	G4	1931-1940; 1941-1950	0.3	0	0.23	0.516	0.607	0.573
Larix kaempferi	Japanese larch	76	Plants	Vascular plants	G4	1971-1980; 1981-1990	0.28	2 04E-47	0.34	0.615	0.707	0.585

Lepidium campestre	Field pepperweed/pepperwort	77	Plants	Vascular plants	J2	1931-1940; 1941-1950	0.58	2 16E-14	0.13	0.367	0.497	0.463
Malva neglecta	Buttonweed cheeseplant, cheeseweed, dwarf mallow or roundleaf mallow	77	Plants	Vascular plants	J4	1931-1940; 1941-1950	0.63	3 19E-08	0.23	0.441	0.603	0.535
Myosotis arvensis	Field forget-me-not	77	Plants	Vascular Plants	J2	1971-1980; 1981-1990	0.57	6 39E-14	0.21	0.434	0.611	0.527
Papaver dubium	Long-headed Poppy	77	Plants	Vascular plants	J4	1971-1980; 1981-1990	0.62	3 45E-33	0.21	0.496	0.555	0.546
Picea abies	Norway spruce	76	Plants	Vascular plants	G4	1951-1960; 1961-1970	0.27	6 43E-54	0.13	0.372	0.503	0.418
Picea glauca	White spruce	76	Plants	Vascular plants	G4	1971-1980; 1981-1990	0.23	8 12E-64	0.19	0.392	0.555	0.445
Picea sitchensis	Sitka spruce	76	Plants	Vascular plants	G4	1971-1980; 1981-1990	0.3	3 56E-90	0.16	0.361	0.455	0.463
Pilosella aurantiaca	Orange hawkweed	77	Plants	Vascular plants	J2	1971-1980; 1981-1990	0.51	1 47E-41	0.13	0.378	0.434	0.376
Pilosella aurantiaca	Fox-and-cubs	77	Plants	Vascular Plants	J2	1971-1980; 1981-1990	0.43	3 93E-64	0.16	0.421	0.528	0.501

Pseudotsuga menziesii	Douglas fir	76	Plants	Vascular plants	G4	1951-1960; 1961-1970	0.25	2 74E-39	0.15	0.443	0.740	0.446
Quercus cerris	Turkey oak	77	Plants	Vascular plants	G1	1931-1940; 1941-1950	0.63	6 72E-06	0.12	0.385	0.482	0.458
Quercus ilex	Holm Oak or Holly Oak	77	Plants	Vascular plants	G1	1931-1940; 1941-1950	0.36	0	0.41	0.623	0.803	0.685
Quercus rubra	Northern Red Oak or Champion Oak	77	Plants	Vascular plants	G1	1971-1980; 1981-1990	0.25	7 44E-17	0.16	0.421	0.830	0.478
Rhus typhina	Staghorn sumac	77	Plants	Vascular plants	G1	1971-1980; 1981-1990	0.23	3 76E-11	0.17	0.453	0.495	0.482
Salix alba	White willow	77	Plants	Vascular plants	C3	1931-1940; 1941-1950	0.67	7 08E-22	0.12	0.386	0.497	0.448
Salix triandra	Almond willow or Almond-leaved willow	77	Plants	Vascular plants	C3	1931-1940; 1941-1950	0.24	6 75E-10	0.14	0.362	0.527	0.479
Salix viminalis	Common osier or osier	77	Plants	Vascular plants	C3	1931-1940; 1941-1950	0.31	1 14E-17	0.17	0.410	0.493	0.412
Sambucus racemosa	Red elderberry	77	Plants	Vascular plants	G1	1971-1980; 1981-1990	0.35	1 94E-56	0.13	0.416	0.474	0.464

Sinapis arvensis	Wild mustard or Charlock	77	Plants	Vascular plants	J4	1931-1940; 1941-1950	0.72	1 09E-87	0.14	0.354	0.770	0.454
Thuja plicata	Western red cedar	76	Plants	Vascular plants	G4	1971-1980; 1981-1990	0.3	1 76E-23	0.15	0.435	0.517	0.400
Tsuga heterophylla	Western hemlock	76	Plants	Vascular plants	G4	1971-1980; 1981-1990	0.3	1 46E-21	0.2	0.466	0.557	0.448
Tsuga heterophylla	Western hemlock-spruce	76	Plants	Vascular plants	G4	1931-1940; 1941-1950	0.27	1 45E-33	0.17	0.395	0.528	0.418

Table 2.2 Results of computer simulations comparing GP to (a) centre of minimum distance (CMD), spatial mean (SMM) and spatial median (SSM), and (b) a kernel density model. The results shown are the mean and SD hit score percentage. Asterisks mark cases in which GP significantly outperformed the other methods. GP performs well across the entire study, never exceeding a mean search efficiency of 4.44% of the simulated area. GP's advantage over other methods increases as the number of sources increases. Other methods fail to replicate this and become increasingly inaccurate as the number of sources increases.

(a) GP versus SSM, SMM and CMD

		Mean hit score % (SD)			
Number of Sources	**Number of Spread Points**	**CMD**	**SSM**	**SMM**	**GP (fitted B)**
1	10	9.5 (0.06)	14.6 (0.08)	19.2 (0.08)	3.9 (0.05) *
1	20	9.4 (0.06)	14.4 (0.06)	19.3 (0.08)	4.3 (0.05) *
1	30	9.2 (0.06)	14.2 (0.07)	19.0 (0.09)	3.3 (0.05) *
2	10	19.4 (4.13)	24.4 (4.44)	29.2 (4.42)	3.4 (0.03) *
2	20	22.6 (4.26)	27.7 (4.15)	32.5 (4.35)	3.6 (0.03) *
2	30	27.1 (4.13)	32.1 (4.15)	36.9 (4.41)	3.5 (0.04) *
5	10	28.0 (4.26)	38.1 (4.49)	42.9 (4.55)	2.7 (0.02) *
5	20	35.9 (4.10)	32.6 (4.35)	47.4 (4.4)	3.1 (0.02) *
5	30	40.3 (4.20)	50.9 (4.20)	52.4 (4.52)	2.3 (0.02) *

(b) GP versus kernel density model

		Mean hit score % (SD)	
Number of Sources	**Number of Spread Points**	**Kernel (fixed h)**	**GP (fixed B)**
1	10	4.9 (0.06)	4.4(0.05)
1	20	4.9 (0.05)	4.3 (0.05)
1	30	4.7 (0.07)	4.2 (0.05)
2	10	5.2 (3.06)	3.5 (0.05) *
2	20	5.5 (3.07)	3.3 (0.05) *
2	30	6.8 (3.10)	3.3 (0.05) *
5	10	8.7 (3.06)	2.7 (0.05) *
5	20	7.7(3.07)	2.7 (0.05) *
5	30	13.5 (3.12)	2.6 (0.05) *

Chapter 3: Theoretical development of geographic profiling

3.1 Abstract

I outline the current problems with the criminal geographic targeting (CGT) algorithm, highlighting the problems with the model's notation, form and application. I introduce a corrected version of the notation used by O'Leary (2009) and discuss the further problems with this approach. I introduce the Bayesian paradigm and show the work done by other authors to restructure GP in Bayesian terms. I introduce the problems with a simple Bayesian model and highlight the problem of multiple sources. I develop a Bayesian Dirichlet process mixture (DPM) model in collaboration with Robert Verity. I solve the problem of identifying multiple sources, even when the number of sources is unknown – a requirement for many biological studies. I present a new, rigorous mathematical and computational method, and show why previous Bayesian methods were outperformed by the empirically-developed CGT algorithm used in criminology, and go on to demonstrate that my new method combines the advantages of both previous methods, using simulations and a real-world example (malaria). My approach provides an increase in search efficiency over other methods and is likely to lead to improved targeting of interventions and more efficient use of resources.

3.2 General introduction

The development of geographic profiling has – understandably – been driven by the need for practical solutions to the problems encountered by law enforcement agencies. O'Leary (O'Leary 2010a, 2010b, 2012) placed GP in a Bayesian

framework, mathematically formalising the problem. However, the model put forward by O'Leary makes the simplifying assumption that all observed data points originate from a single source, and hence performs extremely badly in cases where there are actually multiple sources (see Section 3.7). Thus, despite the mathematical appeal of O'Leary's approach, the CGT algorithm continues to be widely used as a result of its proven track record (Rossmo 2000).

Here, I present a well-defined mathematical approach that unifies existing methods under one framework. Crucially, this method explicitly deals with the issue of multiple sources – a situation typical of biological data sets, but less common in criminology. Under these circumstances, my model outperforms both the CGT algorithm and a simple Bayesian model based on O'Leary (2010a). Further, I develop a computational approach using Markov Chain Monte Carlo (MCMC) methods that extends the technique to large data problems. Finally, I demonstrate the effectiveness of the model using a real-life example of malaria cases in Egypt.

Specifically, I assert that (i) one of the reasons for the CGT algorithm's improved performance relative to the simple Bayesian model lies in its ability to deal with multiple sources; hence by constructing a Bayesian model that incorporates the ability of the CGT algorithm to deal with multiple sources while maintaining the mathematical rigour of the simple Bayesian model, we can outperform both of the existing methods; (ii) this method can be extended to large data problems using MCMC; (iii) this method can be used to provide practical solutions to real-life problems, such as those found in epidemiological applications.

3.3 The CGT model

In this section I will discuss the problems that have been identified in the CGT algorithm from Rossmo (2000). These include both mundane problems such as modern and consistent notation and more serious ones such as mathematical flaws and its incorrect use as a probability estimate. As described in Chapters 1 and 2, the CGT grew out of the background theory of criminology. It was seen as only one analytic tool available to a geographic profiler. When used, it was viewed as being placed within a framework of an investigation with caution and its results were always considered under expert guidance (Rossmo 2000; Laverty 2002). This system was very practical, but led to the end of mathematical development of the CGT algorithm. Adjustment to beliefs of the modeller were not carried out at the model design phase but rather applied to the data or the results after the CGT had been implemented (Laverty 2002). I believe that the model itself needs adjustment to differing conditions and must be developed in a rigorous framework to allow such extensions to exist.

The first problem I encountered was the rather archaic notation of the CGT (Rossmo 2000 and Raine *et al*. 2009). This has one small practical problem as it means that the model is inaccessible to the majority of modern statistics and mathematics literature. In addition, the choice function set out by Rossmo (Figure 1.4) results in some instances in the model being divided by zero, resulting in a non-real number at that location, which is then subsequently replaced by a zero. This is mathematically atrocious; a more elegant formulation of the model would both make it accessible and remove this overt error.

Just as I was deriving a new formulation of the CGT model, O'Leary (2009)

published the first of his manuscripts on Bayesian GP in criminology. Here he addressed the same problem and produced the first published alternative to the CGT written in modern mathematical notation. I strongly agree with the ideas put forward in O'Leary (2009) and have subsequently adopted his notation for the remainder of the thesis. O'Leary defines the CGT as the following:

$$f(d) = \begin{cases} \frac{k}{d^h} & if \quad d > B, \\ \frac{kB^{g-h}}{(2B-d)^g)} & if \quad d \leq B. \end{cases}$$

(Equation 3.1)

where the CGT surface is generated as a function of distance *d* (usually Manhattan or Euclidean) with three parameters *h* and *g* corresponding to Rossmo's *f* and *g* and *B* being exactly the same as Rossmo's *B*. *k* remains a scaling constant for the height of the surface. For any series of spatial points x and y the distance between them in Euclidian space is:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

(Equation 3.2)

Note that I have corrected a mistake made in the original article in the second equation. The inequalities were reversed in the original O'Leary (2009) publication. This version of the model solves the issues of dividing through by zero and also lends itself to modern computer programming, being easily implemented using if

and else functions in C++ or R (R Core Team 2012).

Having established the CGT as a useful algebraic function using modern notation, there still remains a problem, in that the CGT does not deal with probabilities; rather, it produces a ranked surface (discussed in Chapter 2). Rossmo has consistently stated that this is of little concern (Rossmo 2000; Rossmo pers. comm.): law enforcement officers are only interested in a ranked search surface that a probability surface will be transformed into anyway. Again, it is my belief that the benefits offered by using a probabilistic framework are numerous and important. Converting GP into a probabilistic framework as in Section 2.2 offers the benefits of applying the considerable wealth of analytical and theoretical tools developed in modern statistics. These include but are not limited to: true statistical outputs such as likelihoods, hypothesis tests, model averaging and comparison, Bayesian prior information and solving problems analytically using calculus. In addition, putting GP in this rigorous framework will immediately make it more appealing and accessible to researchers outside of criminology.

O'Leary's Bayesian work (2009, 2010a) showed that the maximum likelihood estimate of a normal distribution used in GP always converges on the spatial mean as a single point source estimate. He also discarded the idea of a pure likelihood-based approach and suggested a Bayesian framework would be perfect for GP. This would also have the threefold advantage of maintaining our uncertainty of our parameters, allowing the use of (sometimes detailed) prior information and connecting the model with a developing and active field of research.

3.4 O'Leary's Bayesian model

Here I will produce a brief outline of Bayesian inference sufficient to frame the GP problem and discuss the developments of O'Leary (2009, 2010a) and, to a lesser extent, Canter (see for example Canter *et al.* 2006) who both brought GP into a Bayesian framework with differing degrees of success. I will show some performance estimates of a simple Bayesian GP that I developed based on O'Leary's approach and discuss why I think this simple model can be dramatically improved.

As discussed in Section 1.4 Rossmo's original approach was not phrased in Bayesian terms, but is almost described in Bayesian language. The use of the terms 'invert' and 'inversion' by Rossmo show clearly that there has been a mental jump that incorporates the use of inverse probability, even if this was not carried through mathematically. Bayes' theory is a description of how we can use inverse probability, so that data may inform our hypotheses. The GP problem is inherently a Bayesian one. We wish to use data (crime sites) to inform our hypotheses (location of anchor points of offenders). Bayes' rule can be derived from a simple set of rules taken from basic probability theory.

The probability of an event A given an event B is given by the following:

$$P(B|A) = \frac{P(A \bigcap B)}{P(A)}$$

(Equation 3.3)

Equivalently, the probability of the event B given the event A is given by the following:

$$P(A)|B = \frac{P(A \bigcap B)}{P(B)}$$

(Equation 3.4)

If these two equations are rearranged it is possible to arrive at:

$$P(A|B).P(B) = P(A \cap B) \ = P(B|A).P(A)$$

(Equation 3.5)

Then if the right and left hand sides of the equation above are divided through by P(B) we arrive at Bayes' theorem:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

(Equation 3.6)

Note that it is entirely possible to reverse the symbols A and B as they are arbitrarily chosen. By dividing through by P(A) we would arrive at a statement of Bayes theorem with the two symbols reversed.

In GP it is possible to imagine the probability of crime location being informed by our knowledge of crime locations. The full description of GP in a Bayesian formulation was first published by O'Leary (2009). His main results were the following:

$$P(z, \alpha|x) = \frac{P(x|Z, \alpha)}{P(x)}$$

(Equation 3.7)

Where z is a single stable anchor point, the average offense distance is α and the location of a crime is *x*. From this Bayesian formulation O'Leary is able to arrive at the following:

$$P(z|x_1, ..., x_n) \propto \int P(x_1|z, \alpha)...P(x_n|z, \alpha)H(z)\pi(\alpha)d\alpha)$$

(Equation 3.8)

This is a fundamental statement of the GP problem: the probability of the location of the anchor point (P(z)) is proportional to the integrated probability of the locations of the crime series given an average offence distance and search area (*H*) and prior information on the search area π. For a full derivation of this results see O'Leary (2010).

I compare the CGT algorithm against a simple Bayesian model based on the initial approach described by O'Leary (2010a, 2012), and ignoring subsequent extensions relating to the choice of priors. This approach differs from the CGT in that distributions are defined and manipulated according to the laws of probability. The starting point is to write down the likelihood of the data, given the known location of

the source. This is achieved through the use of a probability distribution, which I will refer to as the migration profile, in which the likelihood of finding an observation at any point in the domain is expressed relative to the location of the source. Assuming independence between observations, the likelihood of the sample is simply the product over the likelihoods of the individual data points (in fact, Rossmo (2000) considered a version of the CGT using summed log space). By placing a suitable prior on source location and applying Bayes' rule it is possible to derive the posterior distribution of the source location, given the observations (O'Leary 2009).

Unsurprisingly, the choice of method makes a big difference to the results. While the CGT algorithm tends to create a patchy distribution of peaks and troughs, entertaining the possibility of a number of different source locations (Rossmo 2000), the simple Bayesian method tends to place all the posterior probability mass around the spatial mean of the data points (O'Leary 2009). Another important difference between the methods is in the rate of convergence. In the Bayesian approach the variance of the posterior distribution decreases rapidly as more data is added, whereas in the CGT method the variance of the geoprofile can never be less than the variance of the decay function. Generally, when there is in fact a single source location the Bayesian method is predicted to outperform the traditional method (O'Leary 2009). However, if there is the potential for multiple source locations then the Bayesian method is predicted to converge quickly on the wrong answer, while the traditional method will still perform well. In this study, we test this prediction using a variety of simulations (see Section 3.7).

3.5 The Dirichlet process mixture model

My primary objective in this section is to address the issue of multiple sources within a well-defined Bayesian framework. The tool that allows us to do this is the Dirichlet Process Mixture (DPM) model, which has a strong mathematical foundation (Ferguson 1983; Green & Richardson 2001) and is finding increasing application within biology (eg Huelsenbeck *et al.* 2006; Huelsenbeck & Andolfatto 2007; Dorazio *et al.* 2008). Unlike many clustering approaches, DPM models do not require the user to specify the number of clusters beforehand, and are therefore extremely useful in situations in which there is no strong prior information about the exact number of clusters. In place of a fixed number of clusters, the DPM model describes the process of cluster formation using a single 'concentration parameter' α. Specifically, if we have already seen n observations, of which n_A came from group A, then the probability of the next observation also belonging to group A is given by $n_A/(n + \alpha)$. It follows that, no matter how many observations we have seen, there is always the probability $\alpha /(n + \alpha)$ of the next observation originating from a previously undiscovered group. While we may not believe there to be a truly unlimited number of groups, by allowing for the possibility of an expanding number of groups we can ensure that our model is always appropriate for the quantity of data at hand. Obviously the choice of the concentration parameter α has a strong influence on the model. Although an appropriate value of α could be fitted from training data, I chose instead to integrate over our uncertainty by placing a diffuse hyper-prior over α. Where stronger prior information is available, the model can easily be adapted to include this.

The DPM model goes on to calculate the probability of the data given a particular grouping. This part is mathematically very similar to the simple Bayesian model,

with the probability of each observation being calculated using the appropriate migration profile centred on the associated source location. By using Bayes' rule and an appropriate prior it is possible to invert the problem, and calculate the posterior distribution of source locations, given the data. A more detailed description of the model, including expressions for the posterior quantities of interest, is as follows:

The Dirichlet process mixture model

Note: This section was written in collaboration with Robert Verity.

Most of the groundwork in this area has already been set out by O'Leary (2009) and Neal (2000), and we follow their notation closely. For the most part we will consider a discrete-space specification of the problem, although results presented here are perfectly generalisable to continuous space.

Thus, let us assume that every crime location and every source location occupies a single cell in a finite grid of cells, which we will call Ω. Every $\omega \in \Omega$ is a vector with two components $\omega = (\omega^{(1)}, \omega^{(2)})$ giving the x-and y-coordinates of the location in two-dimensional space. We presume that we are working with a series of *n* linked crimes, with the crime sites under consideration being labelled $x_1, x_2, \ldots, x_n$. Each x_i is modelled as a draw from the random vector X_i, defined over the sample space Ω. We will also assume a strictly infinite series of potential source locations z_1 to z_∞ (denoted $z_1,\ldots,\infty$), where each z_i is modeled as a draw from the random vector Z_i defined over Ω. The prior on source locations (also the base distribution of the DP) is given by the distribution of Z_i, which we will call G_0.

Each observation x*i* is linked to a particular source location via an associated categorical variable c_i. The value of c_i indexes one of the infinitely many potential

source locations, meaning the observation x_i can be said to have originated from source location z_{ci} . Specifically, we assume that the observation x*i* is a draw from the distribution $F(z_{ci})$, which could, for example, be a distribution centred on the point z_{ci} . The values of the categorical variables c_i are drawn from a Chinese restaurant process with concentration parameter α. Finally, we model α as a draw from the hyper-prior distribution *H*. The complete DP model can be written as follows:

$$x_i | z_{c_i} \sim F(z_{c_i})$$

$$z_{1,\ldots,\infty} \sim G_0$$

$$c_i \sim CRP(\alpha)$$

$$\alpha \sim H.$$

(Equation 3.9)

It is important to note that uppercase italicised letters such as G_0, *F* and *H* describe types of distribution (e.g. normal, exponential), with the associated probability mass/density functions being written $G_0(z_i)$, $F(x_i \mid z_{ci})$ and $H(\alpha)$ respectively.

When considering the problem of posterior inference from this model it is useful to imagine that the categorical variables $c_1 \ldots c_n$ define a particular partition of the data into *s* distinct groups. We can then define our quantities of interest conditional on a particular partition, before eventually summing over all partitions, weighted by their posterior probability.

For example, in most applications the main quantity of interest to us is the Bayesian

analogue of a traditional jeopardy surface. At any point in the grid, this distribution tells us the probability that there exists a realised source (i.e. any source from which at least one of our observations originated). For a particular partition, this is given by the union of the within-group posterior distribution of source location over all s groups. Conversely, the posterior probability of the partition can be obtained (up to a constant of proportionality) by marginalising the likelihood over all source locations.

The DPM model can be adapted to use any type of migration profile; here, we use a bivariate normal distribution with standard deviation σ. Similarly for the prior on source location, we assume a normal distribution with standard deviation τ.
In order to obtain an analytical solution to the DPM model described above we would be required to sum over all possible partitions of the n data points into up to n groups. The number of such partitions is given by the n^{th} Bell number (B_n) which becomes prohibitively large for values as low as n=10 (B_{10}=115,975). Thus, for any reasonably sized dataset we must turn to MCMC methods for a practical solution. Fortunately, a detailed exposition of MCMC algorithms for DPM models is provided by Neal (Neal 2000), and we needed only to adapt these algorithms to our specific needs, as follows.

Details of the MCMC algorithm

Note: This section was written in collaboration with Robert Verity.

The core of our MCMC algorithm is centered on **algorithm 2** in Neal (2000). The algorithm is essentially a Gibbs sampler, which works by alternately drawing new categories and new source locations from their respective conditional distributions.

We start by drawing a new value k for category c_i from its conditional distribution

given our most up-to-date values of the source locations. Thus, in agreement with Neal's original notation, let c_{-i} be the subset of elements c_j for which $j \neq i$, and let $n_{-i,k}$ be the number of c_{-i} that are equal to *k*. Then, adapting Neal's formula 3.6 to our specific needs we can write:

If

$$k \in c_{-i} \quad ,$$

$$P(c_{-i} = k | c_{-i}, x_i, z_i, ..., \infty, \alpha) = b \frac{n_{-i,k}}{n - 1 + \alpha} F(x_i | z_k);$$

If

$$k \notin c_{-i} \quad ,$$

$$P(c_i = k | c_{-i}, x_i, z_i, ..., \infty, \alpha) = b \frac{\alpha}{n - 1 + \alpha} \sum_{\omega \in \Omega} F(x_i | \omega) G_0(\omega),$$

(Equation 3.10)

where *b* is a normalizing constant that ensures that the above probabilities sum to unity. This sampling scheme is repeated for *i =1,...,n*.

Then, for each *i*, we can draw a new value of the source location z_{ci} from its conditional distribution given our most up-to-date values of the categorical variables. This distribution will be made up of the prior G_0 and all observations x_j for which $j \neq i$ and $c_j = c_i$. In fact, we usually do not have to repeat this process for all *i =1,...,n*, as it is only necessary that each unique group is re-sampled at least once. Our main extension to this methodology has been to integrate over the hyper-prior on α. To

this end, let us define the function *t(x)* as follows:

$$t(x) = \int_0^{\infty} \frac{\Gamma(\alpha)\alpha^x}{\Gamma(n+\alpha)} H(\alpha) d\alpha,$$

(Equation 3.11)

where $H(\alpha)$ is our prior over α. This function can be pre-computed for $x = 1,...,n$ outside of the main workings of the sampler. Then, we can re-define (Equation 3.10) with α integrated out as follows:

If

$$k \in c_{-i} \quad ,$$

$$P(c_i = k | c_{-i}, x_i, z_i, ..., \infty, \alpha) = b' t(u_{-i}) n_{-i,k} F(x_i | z_k);$$

If

$$k \notin c_{-i} \quad ,$$

$$P(c_i = k | c_{-i}, x_i, z_i, ..., \infty, \alpha) = b' t(u_{-i} + 1) \sum_{\omega \in \Omega} F(x_i | \omega) G_0(\omega);$$

(Equation 3.12)

where u_{-i} denotes the total number of unique values in c_{-i} (i.e. the total number of groups defined by the categorical variables once element *i* is removed). The notation b' has been used to indicate that this constant of proportionality is not equivalent to *b*

in Equation 3.11.

A more powerful algorithm can be used under the assumption of continuous space and conjugate prior and likelihood. In this case $F()$ and $G_0()$ represent conjugate probability density functions, and the following result obtains:

If

$$k \in c_{-i} \quad ,$$

$$P(c_i = k | c_{-i}, x_i, z_i, ..., \infty, \alpha) = b'' t(u_{-i}) n_{-i,k} \int_{\Omega} F(x_i|\omega) dW_{-i,k}(\omega);$$

If

$$k \notin c_{-i} \quad ,$$

$$P(c_i = k | c_{-i}, x_i, z_i, ..., \infty, \alpha) = b'' t(u_{-i}+1) \int_{\Omega} F(x_i|\omega) dG_0(\omega),$$

(Equation 3.13)

where $W_{-i,k}(\omega)$ is the posterior distribution of ω based on the prior G_0 and all observations x_j for which $j \neq .i$ and $c_j = k$. The notation b'' indicates that this is yet another constant of proportionality. In reality none of these b-values need to be evaluated explicitly, as it is only necessary that we can sample from the requisite distribution. Equation 3.13 can be substituted directly for formula 3.7 in Neal's algorithm 3 (Neal 2000).

3.6 Model implementation

The model is written in R (R core team 2012) and integrates with Google Maps via the R package RgoogleMaps (Loecher 2012). I set τ to the maximum distance in either latitude or longitude between the crime sites. τ equals one standard deviation of the normal distribution centred on the source; hence, around two-thirds of the time the source will lie within this distance of the centre, and the model allows for sources well outside the area bounding the crimes. In some cases, there will be biological data on dispersal patterns that can be used to inform the choice of σ; for example, studies have shown that most malaria transmission occurs close to the larval breeding sites – usually between a few hundred meters and a kilometer – and rarely exceeds 2-3 km (Carter *et al.* 2000). In other instances, there may be little or no data on which to draw, for example with outbreaks of new diseases, or invasions by novel and/or poorly-studied species. In such cases, a good estimate of σ may be obtained by plotting a histogram of pairwise distances between all incident locations, and setting σ to the value of the first peak, since this is likely to correspond to within-cluster distances. My experience suggests that this should be done using all event location data, including duplicates from the same location, but that any zero values in the resulting histogram should be ignored when selecting the resulting peak from the first histogram.

When running the MCMC, multiple chains were run simultaneously, with convergence being assessed using the Gelman-Rubin diagnostic statistic (Gelman *et al.* 2003). After the burn-in period, samples were obtained until the largest standard error for any point on the estimated surface was less than 0.01. Samples were not thinned, as it has previously been shown that this does not increase statistical power (Link & Eaton 2012).

3.7 Methods and results

Comparing the simple Bayesian, CGT and DPM models

I simulated 6, 7, 8 or 9 data points from the migration profile in Equation 3.1 (B = 0.5, f = 4, g = 4), emanating from either 1, 2 or 3 sources in each case. I use a 100*100 grid, and replicated each combination of spread points and sources 1000 times. I used a fully factorial ANOVA to test the effect on the hit score percentage (or average hit score percentage when the number of sources was > 1) of model type, number of sources and number of spread points. Three model types were examined: the analytical form of the DPM model, the classical CGT algorithm (Rossmo 2000) and the simple Bayesian model. To simplify the comparisons between models, both the simple Bayesian and the DPM models were adapted to use the same distribution described in the CGT, rather than the normal distribution described previously. Model type, number of points and number of sources all significantly affected the relative performance of the three models (ANOVA: model type: $F_{2,35964} = 4787.05$, p< 2e-16; sources: $F_{2, 35964} = 13099.30$, p < 2e-16; points: $F_{3, 35964} = 106.23$, p<2e-16). All interactions were highly significant, with the F value for model type*sources interaction indicating that this was the most important of these ($F_{4, 35964} = 2840.12$, p<2e-16); none of the other F values exceeded 52. Tukey post-hoc tests at $\alpha = 0.05$ showed that (i) the CGT significantly outperformed the simple Bayesian model, by an average of 2% (95% CI: 1.75-1.86%); (ii) my model showed a statistically significant improvement over both the CGT algorithm, albeit only by 0.3% (95% CI: 0.25-0.36%) and the simple Bayesian model, again by about 2% (95% CI: 2.1-2.2%). Across all 12,000 runs, the DPM model performed as well or

better than the CGT in 74.9% of trials, and as well or better than the simple Bayesian model in 91.5% of trials. However, although the DPM model outperformed the simple Bayesian model overall, the simple Bayesian model had a small advantage when there was a single source (Figure 3.1).

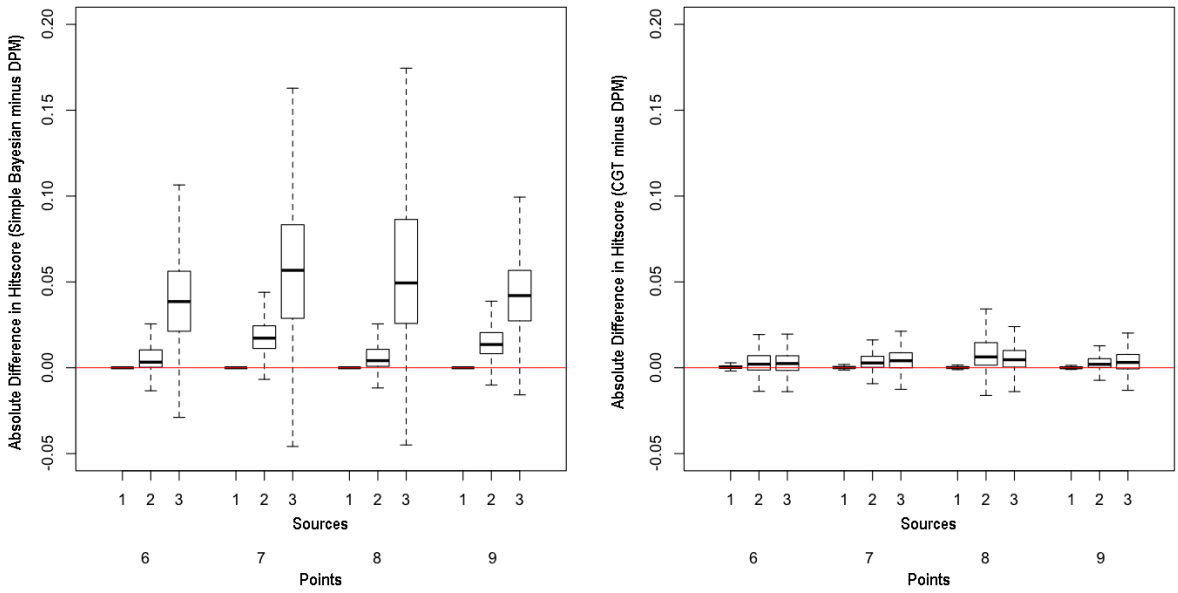


Figure 3.1 Comparison of the analytical form of the DPM model against (left) the simple Bayesian model, and (right) the CGT algorithm, expressed as the hit score percentage of the simple Bayesian model minus the hit score percentage of the DPM model, and the hit score percentage of the CGT algorithm minus the hit score percentage of the DPM model, respectively. Thus, points above the red line indicate cases in which the DPM model outperformed the other models. In both cases, the DPM model has a statistically significant advantage, although this is more pronounced for the comparison with the simple Bayesian model. In both comparisons, the relative performance of the DPM model improves as number of sources increases.

I also repeated the comparison of the DPM model with the CGT for larger data sets (1, 2 and 5 source locations; 20, 40, 60, 80 and 100 spread points), using just the MCMC implementation of the model, with extremely similar results (ANOVA: model type: $F_{1,29992} = 167.7$, p<2e-16; sources: $F_{2,\ 29992} = 10603.1$, p<2e-16; points:

$F_{4, 29992}$ = 1986.2, p<2e-16; model type*sources: $F_{2, 29992}$ = 463.5, p<2e-16; model type*points: $F_{4, 29992}$ = 17.4, p<2e-16; sources*points: $F_{8, 29992}$ = 2916.7, p<2e-16; model type*sources*points: $F_{8, 29992}$ = 0.9, p = 0.87). Tukey post-hoc tests at α = 0.05 showed that my model outperformed the CGT algorithm in a statistically significant way; again, this improvement was most marked when the number of sources was > 1 (Figure 3.2).

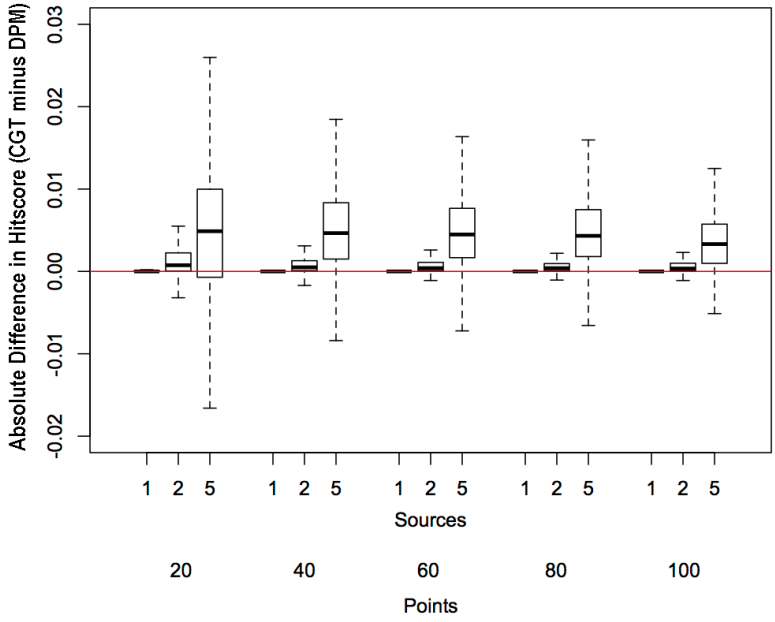


Figure 3.2 Comparison of the MCMC implementation of the DPM model against the CGT algorithm, expressed as the hit score percentage of the CGT algorithm minus the hit score percentage of the DPM model. Again, points above the red line indicate cases in which the DPM model outperformed the other model. The DPM model outperformed the CGT algorithm, especially as number of sources increases.

MCMC validation

For the reasons described above, the analytical form of my model can deal with only small datasets, and for larger datasets an MCMC implementation is required. For each of the 12000 simulations described in the first part of the Results (1000

replicates of each combination of 1, 2 and 3 sources and 6, 7, 8 or 9 spread points), I also used the MCMC algorithm described above and calculated the correlation between the surface produced by the analytical form of our model and the MCMC form. These were extremely highly correlated (r (mean ± sd) = 0.9998 ± 0.0010), demonstrating that the MCMC algorithm does indeed find the same – or at least extremely similar – posterior distributions as the analytical form of the model.

Case study

I tested the performance of our model in a real world example by using the MCMC implementation of the DPM model to reanalyse data from Le Comber *et al.* (2011). In this study, spatial data relating to 139 recorded *Plasmodium vivax* malaria cases were collected, and buffer zones of 2 km were created around the locations of these malaria cases and merged to form a polygon of 296.5 km^2 (Hassan 2006). All accessible aquatic habitats within this study area (surface/cryptic; temporary/semipermanent/permanent) were located and characterised between April and September 2005. These included water tanks, water pools created through pipelines or drainage system breakage, seepage from slum housing, natural springs, pools and ditches filled with ground water. Water sources included in this analysis were identified as bodies of water harbouring at least one mosquito larva over the study period (n=59). A total of 11 mosquito species were identified, including the malaria vectors *Anopheles sergentii* and *Anopheles pharoensis*, as well as other, non-vector, species. Of these 59 sites, seven tested positive for one or both of the malaria vectors *Anopheles sergentii* and *Anopheles pharoensis* (*Anopheles sergentii* is well established as the most dangerous malaria vector in Egypt (El Said *et al.* 1986)).

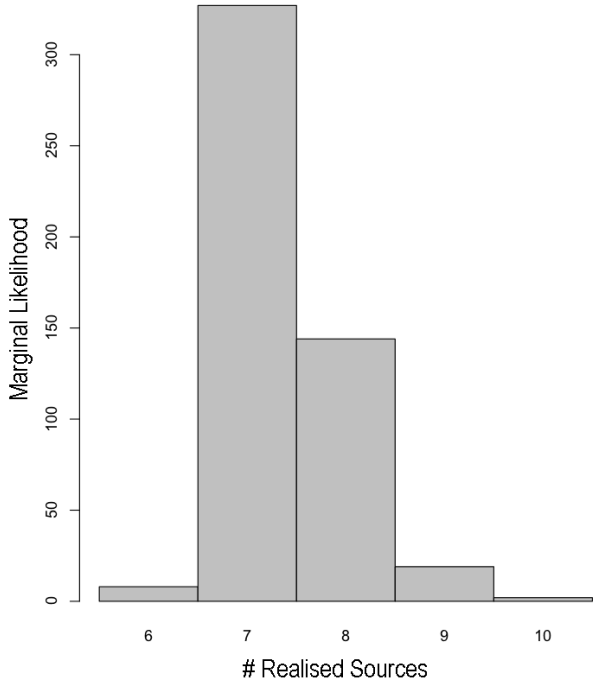


Figure 3.3 Marginal likelihood of different numbers of realised infection sources for the Cairo data. The DPM model estimates that there are 6-10 sources, and assigns the highest likelihood to seven sources.

The median hit score percentages for the seven vector breeding sites identified in Hassan (2006) were 0.34% for the DPM model, compared to 0.43% for the CGT and 1.2% for the simple Bayesian model (note that the hit score percentages for CGT differ from those previously published in Le Comber *et al.* (2011) since the R implementation of the CGT uses a slightly different search area dictated by Google Maps, and the denominator in the hit score calculations thus differs). For five of the seven sites, hit score percentages for the DPM were less than half a per cent. An additional output of my model is a barplot of the posterior probability of the number of realised sources (Figure 3.3). My model indicated the highest likelihood for seven sources, with a likely range of 6-10. Interestingly, these were not all in the same locations as those identified in the original study (Figure 3.4). One possibility, of

course, is that the model is wrong; on the other hand, if the model is correct, this would suggest that some sources were missed in the original survey – not unlikely, given the difficulty of locating small, transient breeding populations of mosquitoes.

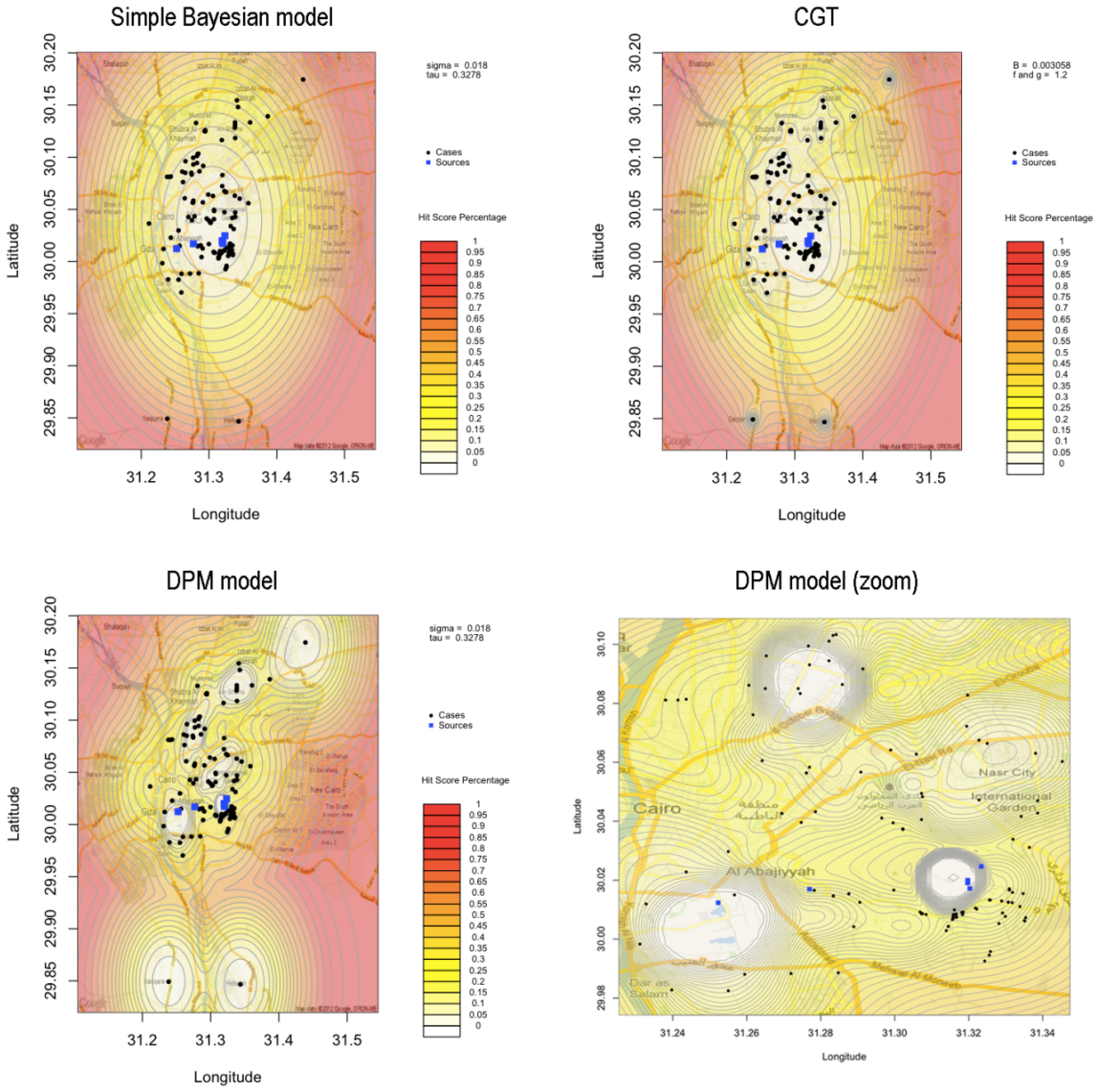


Figure 3.4 Geoprofile from 139 *Plasmodium vivax* cases (black dots) in Cairo, Egypt, using (top left) the simple Bayesian model; (top right) the CGT algorithm; (bottom left) the DPM model. A close-up of the DPM surface is shown in the bottom right.

3.8 Discussion

My model improves on both the CGT and the simple Bayesian model, retaining both the practical utility of the CGT algorithm and the mathematical rigour of the Bayesian approach. In simulations, it outperforms the CGT algorithm, in a small but statistically significant way when there is a single source, and to a greater extent when there are multiple sources. This chapter shows that the DPM model performs well in both rigorous simulations and in a real-world example.

In its construction, the DPM model forms a bridge between the seemingly disparate methodologies of the CGT and the simple Bayesian approach to geographic profiling. From a practical point of view the major difference between the two existing approaches lies in whether distributions should be summed (CGT) or multiplied (simple Bayesian). The DPM model works by splitting the data into groups, with each group corresponding to a different source location. The laws of probability then dictate that distributions should be multiplied within groups, but summed between groups. Thus, if all points are assigned to a single source we arrive back at the simple Bayesian model, while if all points are assigned to different sources we arrive at something more akin to the CGT algorithm. In many cases the best fit to the data will lie somewhere in between these two extremes. In this context, the concentration parameter α can be understood as a prior over the complete spectrum of models, which allows us to transition between a single-source model and a multiple-source model. When α is set to zero, the DPM model becomes mathematically equivalent to the simple Bayesian model; conversely, as α tends to infinity, we converge on the CGT algorithm. In the majority of cases – particularly those dealing with biological data – the most likely explanation for the data will often lie between these two extremes. For example, in the malaria analysis, the DPM

model assigned the highest likelihood to seven sources from 139 disease case locations (Figure 3.3).

In my simulations, the DPM model outperformed both other approaches. In cases with a single source – a common scenario in criminology – the improvement, although statistically significant, was minimal. However, formulating the problem in a rigorous Bayesian framework allows a number of useful extensions. First, my model produces a true probability surface, allowing us to calculate the marginal probability of different numbers of sources, as in Figure 3.3. Further, we can produce a probability surface conditional on a particular number of sources, thereby allowing us to break the overall picture down into different scenarios. Second, the DPM model explicitly calculates the posterior probability under the model that a particular crime site is derived from a particular source. This may be of interest in criminology, where crime linkage is an important problem (Rossmo 2000), and may also be useful in biological data sets, where the spatial linkage can be validated against other forms of information (for example genetic data).

So far, the DPM model is constructed with flexibility in mind, rather than statistical power. It is therefore striking just how well the model performs. For particular cases, it will be easily possible to increase the power of the model by incorporation of stronger prior information – for example, inferring the concentration parameter from training data. Similarly, where empirical evidence has shown that non-normal dispersal profiles are appropriate (for example, Cauchy distributions in some bird species (Winkler *et al.* 2005; Van Houtan *et al.* 2007) or bivariate Student's t-distributions in seeds (Nathan & Muller-Landau 2000), these can be used within the same general framework.

One of the most exciting possible extensions of this approach is the utilisation of the

outputs produced by niche models to generate priors in the DPM model. Niche modelling is a well-developed field that has recently been placed on a Bayesian footing (Elith & Leatherwick 2009), making its incorporation into the DPM model relatively straightforward. A Bayesian niche model will produce a probabilistic estimate of the suitability of habitat for the organism being studied that can be used as a prior in the DPM model. Combining these two approaches would go some way towards producing a spatially explicit niche model approach as called for by Peterson (2003).

In epidemiology and invasion biology, much more attention is paid to models that run forwards in time than those that run backwards to locate sources. GP, on the other hand, is radically different, running backwards in time to use current locations to infer sources (Le Comber & Stevenson 2012). However, as O'Leary (2010a, 2012) has shown, a fully Bayesian implementation of GP can easily be extended to run forwards in time. Despite the difficulties faced by all predictive models, this could potentially be important in areas of biology including epidemiology, invasion biology and in conservation biology (eg planning reintroductions of animals or plants).

The DPM model we present here is a general method that can be applied to data describing spread from common source. Evidence-based targeting of interventions is a crucial component in the fight against infectious disease, and targeted interventions are more efficient and more cost-effective than untargeted interventions; for example, malaria is strongly dependent on the location of vector breeding sites, and most transmission only occurs within short distances of these sites (Carter *et al.* 2000). Because of this clustering, untargeted intervention is highly inefficient. In the Cairo study, the DPM model identified five of the seven breeding sites in less than

half a percent of the total search area, representing a dramatic improvement over a non-targeted search.

Although my implementation of the DPM model can deal with large data sets (>>1000 data points), GP methods also work well with very small data sets (Rossmo 2000; Stevenson *et al.* 2012), allowing their use in the early stages of an outbreak or invasion, when control efforts are most likely to be successful. The DPM model provides a useful practical tool for conservation biologists and epidemiologists, offering improvements over other methods that are likely to lead to improved targeting of interventions, and more efficient use of resources.

Chapter 4: Improvements to the DPM model

4.1 Abstract

In this chapter I modify the version of the DPM model described in Chapter 3 so that it fits the parameter σ directly from the point pattern data. In Section 4.2 I first realise the group locations that were originally integrated over in the original DPM model, then I combine this information with the group assignment probabilities. Finally, I create a conjugate inverse gamma prior for σ^2 and arrive at a distribution from which σ can be updated. In Section 4.3 I test this new σ fitting DPM model using simulations and show that the new version of the DPM significantly outperforms the CGT across a range of more realistic distributions. In Section 4.4 I summarise and discuss the implications of these results.

4.2 Fitting σ from point pattern data

The approach to fitting σ described in Chapter 3, where it was estimated from the shape of the histogram of pairwise distances, is unsatisfactory, since it introduces an element of subjectivity. Instead we would like to use the Gibbs update step to allow us to simultaneously fit the clustering, the mean μ and also our parameter σ. I use the same assumptions as in Chapter 3 and start with the simple case of a single normally distributed data with variance σ^2.

Our data is a vector of x that runs from x_1 to x_n

$$\vec{x} = x_1, x_2, x_3 \ldots x_n$$

(Equation 4.1)

We assume that data are drawn from a normal distribution with mean μ and variance σ

$$x_i \sim Normal(\mu, \sigma^2)$$

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

(Equation 4.2)

$$f(\vec{x}|\mu,\sigma^2) = \sum_{i=1}^{n} f(x_i|\mu,\sigma^2)$$

(Equation 4.3)

$$f(\vec{x}|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left\{\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

(Equation 4.4)

The above case shows the basic idea when applied for a univariate distribution. In our case we will be using the bivariate normal, as the data are two-dimensional. The following details the similar steps for the bivariate normal distribution. In this case we assume that there is no correlation between the variances of the two distributions.

This simplifying step allows us to avoid matrix algebra and keep the number of parameters being fitted relatively small (two σ, two μ only).

We can then multiply through by our prior on source location using d_x and d_y as our prior means for each source location in each dimension and T^2_x and T^2_y as our prior variances:

$$f(\vec{x}, \mu_x | d_x, \sigma_x^2, T_x^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \frac{1}{\sqrt{2\pi T_x^2}} exp\left\{ -\frac{1}{2} \left[\sum \frac{(x_i - \mu_x)^2}{\sigma^2} + \frac{(\mu_x - d_x)}{T_x^2} \right] \right\}$$

(Equation 4.5)

This can be rearranged to give the posterior distribution of a source location given the data and the prior.

$$f(\mu_x | \vec{x}, d_x, \sigma_x^2, T_x^2) = \frac{1}{(2\pi\varepsilon_x^2)} exp\left\{ \frac{(\mu_x - \theta_x)^2}{2\varepsilon_x^2} \right\}$$

where

$$\varepsilon_x^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{T_x^2}}$$

and

$$\theta_x = \frac{\frac{\sum x_i}{\sigma^2} + \frac{d_x}{T_x^2}}{\frac{n}{\sigma^2} + \frac{1}{T_x^2}}$$

(Equation 4.6)

Interestingly, when we have an uninformative prior (i.e. as T^2_x tends to ∞) the posterior mean tends to become $\frac{\sum x_i}{n}$, the exact sample mean of the data, while the standard error tends towards $\frac{\sigma^2}{n}$. These are the identical values that we would obtain in a frequentist analysis. This means we can obtain new means for each group by drawing from:

$$\mu_x = Normal\left(\frac{\frac{\sum x_i}{\sigma^2} + \frac{d_x}{T_x^2}}{\frac{n}{\sigma^2} + \frac{1}{T_x^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{T_x^2}}\right)$$

and similarly for μ_y:

$$\mu_y = Normal\left(\frac{\frac{\sum y_i}{\sigma^2} + \frac{d_y}{T_y^2}}{\frac{n}{\sigma^2} + \frac{1}{T_y^2}}, \frac{1}{\frac{y}{\sigma^2} + \frac{1}{T_y^2}}\right)$$

(Equation 4.7)

We can incorporate this information with our group allocation probability. This is discussed in Section 3.5. The probability for the formation of our cluster groups is given by a Chinese Restaurant Process (CRP), integrated over our prior on α. This group formation probability remains the same in this version of the DPM model, expect that previously I had integrated over the distribution of μ_x and μ_y. This is now not possible as I have realised these two parameters. If we let *j* be the datapoint that is temporarily removed from the sample then the vector of probabilities of group assignment is given by:

$$CRP \times \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(x_j - \mu_x)}{2\sigma^2}\right\} \times \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y_j - \mu_y)}{2\sigma^2}\right\}$$

(Equation 4.8)

In order to arrive at an algorithm that will fit both μ and σ from the data it is also necessary to draw a new σ given all the data from all the assigned groups. Let the vector $\vec{C}$describe the allocation of points to groups, such that the ith element C_i gives the grouping of the i^{th} point. We can describe, for example, all the x values in the first group of $\vec{C}$ by using $x_i : C_i = 1$. This reads, "the x_i for which C_i equals 1". Let there be n_j points in the j^{th} group, and let u^j_x and u^j_y be the means in each dimension of the j^{th} group. The probability of the data ion the j^{th} group is given by:

$$f(x_i : C_i = j|\mu_x^j, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n_j}{2}}} exp\left\{ - \sum_{i:C_i=j} \frac{(x_i - \mu_x^j)}{2\sigma^2} \right\}$$

and similarly for $y_i : c_i = j$:

$$f(y_i : C_i = j|\mu_y^j, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n_j}{2}}} exp\left\{ - \sum_{i:C_i=j} \frac{(y_i - \mu_y^j)}{2\sigma^2} \right\}$$

(Equation 4.9)

If there are *k* groups in total the complete probability of the data can be given as:

$$f(\vec{x}|\mu_x^1, ..., \mu_x^k, \sigma^2 = \frac{1}{(2\pi\sigma^2)^{\frac{n_1}{2}}} ... \frac{1}{(2\pi\sigma^2)^{\frac{n_k}{2}}} exp\left\{ -\frac{1}{2} \sum_{i:C_i=1} \frac{(x_i - \mu_x^1)^2}{\sigma^2}, ..., -\frac{1}{2} \sum_{i:C_i=k} \frac{(x_i - \mu_x^k)^2}{\sigma^2} \right\}$$

(Equation 4.10)

There is an analogous expression in the y dimension. The conjugate prior for a normal distribution is the inverse gamma. I set the prior over σ^2 as the inverse gamma. With two parameters delta and beta:

$$\sigma^2 \sim Inverse\,Gamma(\delta, \beta)$$

$$f(\sigma^2|\delta,\beta) = \frac{\beta^\delta}{\Gamma(\delta)}\left(\frac{1}{\sigma^2}\right)^{\delta+1} exp\left\{\frac{-\beta}{\sigma^2}\right\}$$

(Equation 4.11)

We then multiply through by the beta prior and arrive at the model of *x, y* and σ given the parameters μ, δ and β:

$$f(\overrightarrow{x}, \sigma^2|\mu_x^1, ..., \mu_x^k, \mu_y^1, ..., \mu_y^k) = \frac{\beta^\delta}{\Gamma(\delta)}\frac{1}{(2\pi)^n}\left(\frac{1}{\sigma^2}\right)^{\delta+n+1},$$

$$exp\left\{\frac{1}{\sigma^2}\left[\beta + \frac{1}{2}\sum_{i:C_i=1}\left[(x_i-\mu_x^1)^2 + (y_i-\mu_y^1)^2\right], ..., \frac{1}{2}\sum_{i:C_i=k}\left[(x_i-\mu_x^k)^2 + (y_i-\mu_y^k)^2\right]\right]\right\}$$

(Equation 4.12)

This equation is still of the form of the inverse gamma thus we can arrive at our fundamental result:

$$\sigma^2 \sim InverseGamma\left(\delta + n + 1, \left[\beta + \frac{1}{2}\sum_{i:C_i=1}\left[(x_i-\mu_x^1)^2 + (y_i-\mu_y^1)^2\right], ..., \frac{1}{2}\sum_{i:C_i=k}\left[(x_i-\mu_x^k)^2 + (y_i-\mu_y^k)^2\right]\right]\right)$$

(Equation 4.13)

meaning we can draw a new value of σ from an inverse gamma with the parameters above. In the case of this model it is still necessary to determine a prior of the inverse gamma parameters. This can be done in two ways, either using an informative prior (as in Chapter 3 where I knew the mean dispersal distances of mosquitos from previous studies (Carter *et al.* 2000)) or setting the parameters delta and beta to give a diffuse fat-tailed inverse gamma that will offer very little information to the model.

4.3 Simulations to compare different geographic profiling models

Simulation design

I ran simulations to test the two main geographic profiling models: (i) the new DPM model that estimates both μ and σ as well as the clustering pattern from the data directly, using a diffuse inverse gamma prior on σ (Figure 4.1); (ii) the Rossmo model with fixed *f* and *g* and *B* determined from the half nearest neighbour distance.

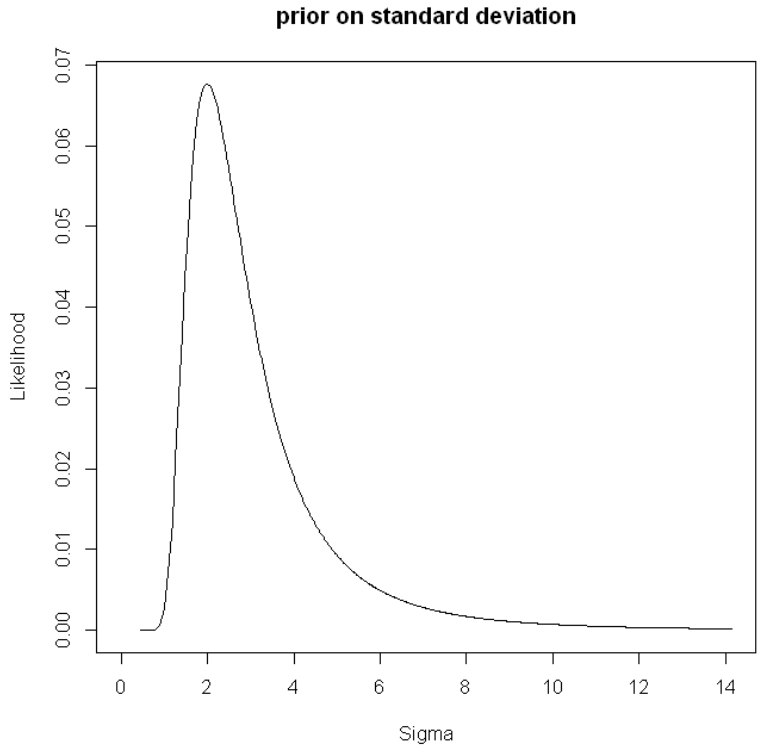


Figure 4.1 Diffuse and weakly informative inverse gamma prior on σ. The expectation was set to five in the 200 by 200 grid of the simulation. Note the low likelihood value assigned to the prior and the fat tail, allowing a wide variety of possible σ values to be explored.

The simulated data are supposed to represent more realistic examples of what data will look like in real-world biological examples. I generated the number of sources from a chi-squared distribution with concentration parameter = 6. From each of these sources I generated N data points from each source where N is a chi-squared distribution bounded to be above 1 with a concentration parameter of 5. The source locations were drawn from a uniform distribution over a 150 by 150 area, and the search area was a 200 by 200 grid (to ensure sources didn't produce points outside the 200 by 200 grid (see below). Four different dispersal distributions were used to draw the sources. These were (i) a normal distribution with standard deviation drawn

from a uniform bounded prior (as Equation 4.5); (ii) a negative exponential distribution with a rate parameter drawn from a uniform bounded prior; (iii) a Cauchy distribution with a location and scale parameter drawn from a bounded uniform prior (Equation 4.14); and (iv) a Student's t distribution with two parameters a mean both drawn from a bounded uniform prior and degrees of freedom set to two (Equation 4.15).

$$f(x, y|x_0, y_0, \gamma) = \frac{1}{2\pi}\left[\frac{\gamma}{((x - x_0)^2 + (y - y_0)^2 + \gamma^2)^{1.5}}\right]$$

(Equation 4.14)

$$f(x, y|x_0, y_0, \nu) = \frac{\Gamma\left[(V + p)/2)\right]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\sum|^{1/2}[1 + 1/\nu(x - \mu)^T \sum^{-1}(y - \mu)]^{(\nu+p)/2}}$$

(Equation 4.15)

In the case of the Cauchy and Student's t the very extremes of the fat-tailed distributions were controlled for in two ways. First, the bounding on the prior prevented the Student's t from becoming especially fat tailed; second, if samples fell outside the search area they were discarded and resampled. Each variant of the simulation was repeated 50 times. Four model types and 50 replicates produced a total of 200 distributions of sources/points, each of which was was analysed by both models.

Results

The new DPM model with an inverse gamma prior on σ outperformed the Rossmo model across a range of distributions (mean (±sd) hit scores: DPM 0.0568 ± 0.0765; CGT 0.137 ± 0.14) (Figure 4.2). The two models were also found to vary in effectiveness in a similar way when presented with different distributions and when presented with large numbers of points and sources. The best fitting model (AIC = -661) used distribution type and method as factors and the number of sources as covariates; including number of points did not improve the model (ANOVA: method: $F_{1,394}$=58.532, p=1.55e-13; distribution $F_{3,394}$=20.520, p=2.25e-12; source number $F_{1,394}$ = 8.059, p = 0.00476) (Figure 4.3).

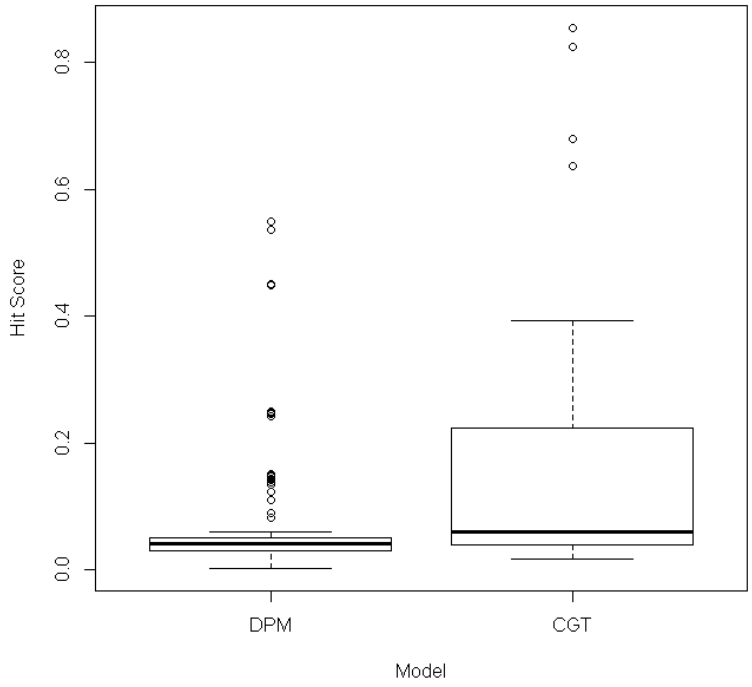


Figure 4.2 The DPM model outperforms the CGT in the simulations. Note the much higher variance of the CGT model.

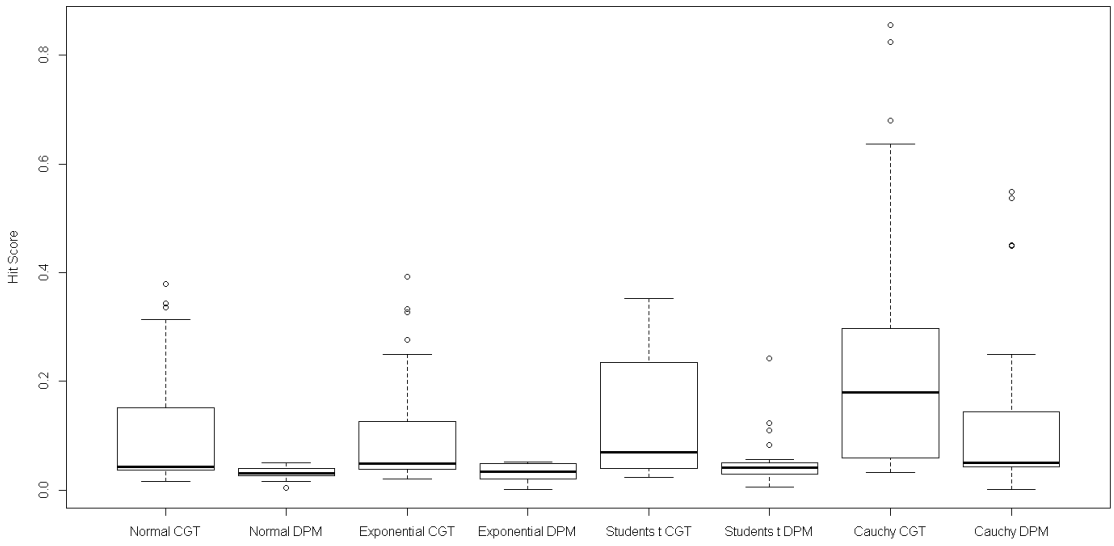


Figure 4.3 The different distributions that the two models were tested against are shown. The DPM performs better across all the different distributions, but both models do worse when presented with Cauchy distributed data and the CGT performs less well when presented with Student's t data.

Discussion of simulation results

The new version of the DPM significantly outperforms the CGT across a wide range of more realistic simulated data. Furthermore, the new model does this using no information except the diffuse prior, itself quickly subsumed by the data, meaning that resultant estimate of the posterior on σ is informed nearly entirely by the data; a similar example of the difference between prior and posterior distribution of σ is shown in Chapter 6 (Figure 6.1).

It is also interesting to note that the choice of distribution does make a significant difference. As Figure 4.3 shows, it is clear that both models perform less well on Cauchy distributed data and to a lesser extent Student's t distributed data. This is of vital importance as it makes it necessary to correctly estimate the form of biological

dispersal for the model to be applied with the greatest efficacy in a real situation (see Chapter 5). Both models perform well when tested against exponential and normal distributions, suggesting that it is possible to apply the model using normal decay functions to an exponentially distributed data set. It is still important to note that the CGT is generally performing well, but is sometimes dramatically failing to locate individual sources with small number of points from a multi-source problem. The new σ fitting DPM does not have this problem and is exceedingly effective at detecting these separate sources with few data points emerging from them.

4.4 Conclusion

The work in this chapter extends the version of the DPM described in Chapter 3 to estimate σ directly from the point pattern data, removing the need for a subjective interpretation of pairwise distance data, although where there is a strong prior on dispersal functions these can be used to make the prior more informative. This new version of the DPM model significantly outperforms the CGT, although the efficiency of both models depends to some extent on the underlying dispersal distribution, underlining the importance of the points discussed in Chapter 5. The complete code for the new version of the DPM model can be found in Appendix D.

Chapter 5: Biological dispersal

5.1 Abstract

The analysis of dispersal has profound consequences in disciplines including population genetics, mate choice, population viability analysis and, importantly, geographic profiling. Unfortunately, the presentation of dispersal data frequently omits crucial information; worse, a simple geometrical mistake commonly leads to incorrect biological inferences. I describe the mathematics of this error and show how to avoid it. I outline three rules for the presentation of dispersal data: (i) Present data as maps, not histograms or single-number metrics alone; (ii) If histograms are used, data should be correctly transformed, with appropriate error bars; (iii) Treat inferences drawn from small datasets with caution. These rules will help authors present data in the most informative way, and ensure that inferences are accurate.

5.2 Spatial transformation and implications for fitting dispersal distributions

The criminal geographic targeting (CGT) algorithm works by assuming that the probability distribution around each crime site has the same form as the criminal's dispersal profile. In criminology, this is almost always the Rossmo distribution (Figure 1.3) (O'Leary (2009) uses the normal distribution, but his method is not often used in practice). However, even in criminology the Rossmo distribution, and in particular the buffer zone, are controversial, and in biology other distributions may be more appropriate (for example, Cauchy distributions in some bird species (Winkler *et al.* 2005; Van Houtan *et al.* 2007) or bivariate Student's t-distributions in

seeds (Nathan & Muller-Landau 2000). In fact, the CGT is extremely robust to the use of different distributions, and works well with biological datasets even with the Rossmo distribution (Le Comber *et al.* 2006; Stevenson *et al.* 2012, Le Comber & Stevenson 2012; Papini *et al.* 2013). Despite this, though, the results of the simulations in Chapter 4 show that significant improvement to GP models may be gained by using the correct dispersal profile, and it is this that this chapter seeks to address (O'Leary 2010a).

5.3 Presenting dispersal data

The study of dispersal is important in fields including population ecology, behavioural ecology, evolutionary biology, epidemiology, invasion biology and conservation (Kot *et al.* 1996; Clobert *et al.* 2001; Nathan 2001; Trakhenbrot *et al.* 2005) and encompasses all major groups of organisms and all types of ecosystems. Unfortunately, dispersal data is difficult to collect (for example Stoner (1992) describes one study in which SCUBA divers followed 259 individual ascidian larvae for up to 15 minutes each as they dispersed) and there is no clear consensus on how it should be reported. Data are sometimes presented as a single number – for example, mean (Reed *et al.* 1988), median (Narbona *et al.* 2005) or maximum (Slough 1989) dispersal distance, while other papers describe distributions of dispersal distances for a particular organism, typically in the form of histograms or fitted models (Beer 1955; Alonso *et al.* 1988; Forsman *et al.* 2002). The most complete descriptions of dispersal distributions are maps that show presence (and potentially absence) over time, but – perhaps for reasons of space in published papers, or perhaps because researchers are unwilling to release detailed data that may

have taken years to collect – this is very often not the case (Table 5.1).

The simplest – and least informative – way in which dispersal data can be presented is as a single number. Usually this is in the form of an estimate of central tendency (mean and/or median), but sometimes the maximum distance travelled by dispersers is presented instead. Even with appropriate measures of variability – for example standard deviation, standard error or ranges – this omits important information about the shape of the underlying distribution.

Table 5.1 Type of data presented in a sample of 142 papers dealing with animal or plant dispersal. Papers were found by searching Google scholar & Web of Science using the keywords; 'Dispersal' & 'Biological Dispersal'. Papers without fully described methods or lacking the raw data for analysis were rejected. Papers counted as presenting map data have been counted if any dispersal maps or fitted models are presented at all, even if some of the data are not presented in this way; the same criterion has been used to count papers presenting histograms. For a full version of this table, including references and histograms, see Appendix C.

Taxon (number of papers)	**Map/Model**	**Histogram (no. correctly transformed)**	**Single number metric only**
Birds (32)	3	21 (1)	8
Mammals (49)	4	15 (2)	30
Plants (35)	15	9 (4)	11
Invertebrates (26)	1	6 (1)	18
Total (142)	*13*	*52 (8)*	*77*

More useful are histograms or fitted models that describe the entire dispersal distribution of the organism in question – and, as more and more authors begin to use custom-built Bayesian tests, the importance of complete distributions for priors or parameter estimation will continue to grow. Unfortunately – as I shall show – a

simple geometrical mistake when data are transformed from a two-dimensional Euclidean space (such as a map) into one-dimensional linear space (for example, a histogram) (see Section 5.4) means that in many published examples such data are misleading, and the biological inferences based on them incorrect (see Section 5.5).

As noted, maps are the best way to present dispersal data, and their use can help avoid some of the errors we discuss. However, they are far from ubiquitous in the published literature. Here, I outline three fundamental rules for the correct use of dispersal data in ecology: (i) Data should be presented as maps, not histograms or single-number metrics; (ii) If histograms are used, data should be correctly transformed, with appropriate error bars; (iii) Biological inferences drawn from small data sets should be treated with caution.

Rule 1: Data should be presented as maps, not histograms or single-number metrics

Some papers provide their data in the form of maps, often using GIS (Winkler *et al.* 2005; Belthoff & Ritchison 2003; Caswell *et al.* 2003; Nicholson *et al.* 2007; Zimmerman *et al.* 2005; Venable *et al.* 2008; Logan & Sweanor 2000). These authors often show direction and distance of dispersal on the map using detailed tracking schematics or with simple direction of travel indicators. The best possible presentation of this data is with a complete point map and complete tracking data. Where this is not possible many authors present release locations and recapture locations on a map. When genetic data is present many authors also present bubble plots or wheel plots. These allow the inference of not just distance of travel but direction of travel too. Using these approaches the maximum possible amount of information is retained, and data presented in this way can be analysed using a

variety of methods, including simple spatial statistics (Gatrell *et al.* 1996; Cressie 1991), kernel density maps (Diggle 1985), species distribution models or geographic profiling (Stevenson *et al.* 2012). Figure 5.1 shows two examples of well formatted maps, one with a simple release and direction grid (Belthoff & Ritchison 1989) and one with a complex fitted model (Clark *et al.* 1999).

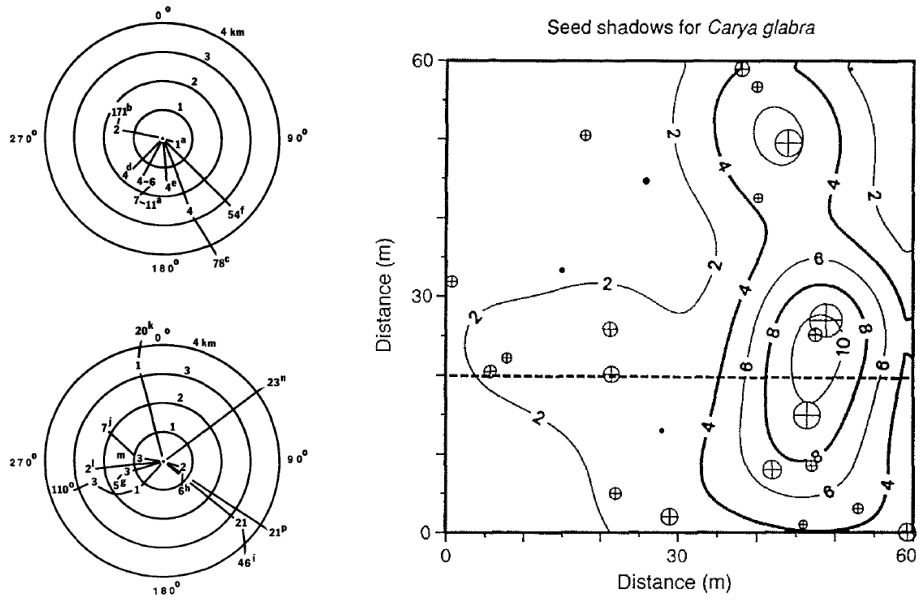


Figure 5.1 Two different types of complete dispersal maps. Left: A simple representation of screech owl dispersal taken from Belthoff & Ritchison (1989). The images clearly show the direction and distance of dispersal for each juvenile owl. Right: Density estimates of seed dispersal from multiple *Carya glabra* trees from Clark *et al.* (1999). The model used contains correctly used two-dimensional data and weights the seed contribution from a number of trees to generate a total seed shadow over the area

Rule 2: If histograms are used, data should be correctly transformed, with appropriate error bars

When dispersal data are presented in linear distributions or as histograms information relating to the angle of deviation from the source has been lost. It is not

possible to recover this information, whatever back transformation is used; it is for this reason that Rule 1 (above) suggests presenting data as maps instead. Nevertheless, histograms can show useful aspects of dispersal and are easy to produce and read. However, there is a second potential difficulty in that the data have been transformed in the process.

This is obvious if we visualise what our histogram looks like in bivariate space. By splitting the data into bins based on Euclidian distance we are essentially slicing the bivariate distribution into concentric rings of constant width, as in Figure 5.2. Every point within the central circle will fall within the first bin of the histogram; every point in the next ring will fall within the second bin of the histogram, and so on. Obviously, since the area of these concentric rings increases with distance from the source, the outer rings will tend to contain more points even when the distribution of points within the study area is uniform. As an example of the problem of visualising a dispersal distribution as a histogram, consider the example of an organism with a symmetrical bivariate normal dispersal distribution (Figure 5.3a); say the density of fruits falling around a tree. It might seem natural to assume that if we plotted the distance of fruits from the tree, we would obtain a histogram like the one-dimensional normal distribution in Figure 5.3b.

This assumption is wrong: instead the curve would follow the Rayleigh distribution – which can be misinterpreted as showing a low density of fruits near the tree. In fact, the dip in the distribution near the origin comes solely from the fact that there are relatively few possible locations this close to the origin. The Rayleigh distribution is obtained by weighting the normal density, at each radius, by the circumference of the circle on which the fruits might fall.

If instead we plotted a histogram of the density of fruits from a transect running out from the centre of the tree, we would have obtained the more readily interpreted normal curve – but then we would only have used a subset of our data. What we want is a way of taking the full set of distances from the tree and obtaining a comparable curve. The situation is more complex when distance decay is added to the data (when organisms are less likely to disperse long distances), but clearly the increased area of the sampling bins still needs to be taken into account. However, as Table 5.1 shows, this is often not the case.

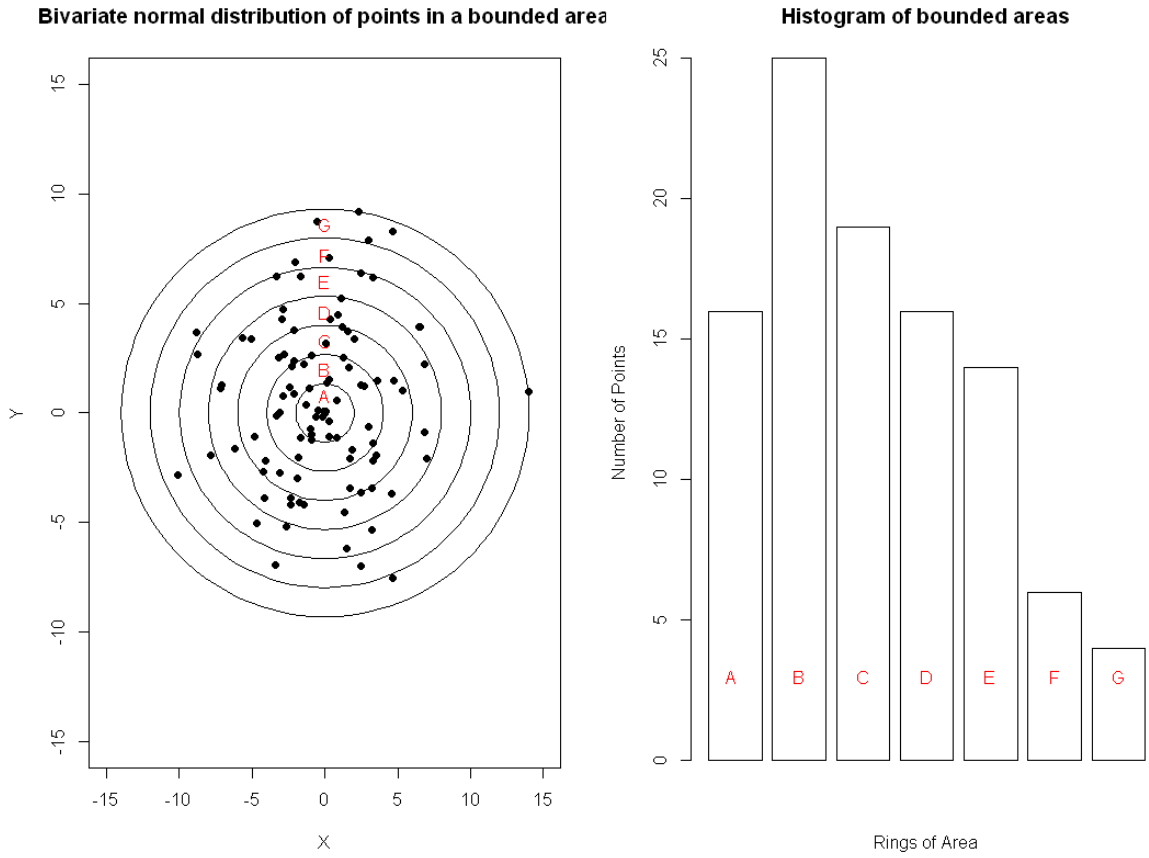


Figure 5.2 Bivariate distribution broken down into a series of bins of increasing area. Even though data are normally distributed in two-dimensions, the increasing area of successive rings of equal width means that abundance will appear artificially inflated as distance from the source increases.

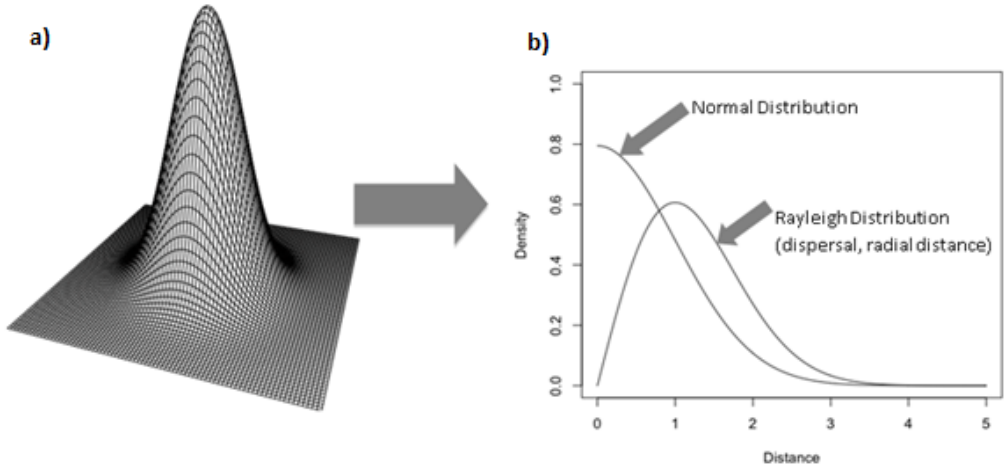


Figure 5.3 Transformations of a normal dispersal kernel a) Shows the a normal density plot, for

example apples falling from a tree

b) Shows the two different ways of plotting this data in two dimensions the Rayleigh distribution (open to misinterpretation), the other the more classic normal (when transformed correctly).

The correct shape of the histogram can be recovered by using an appropriate back-transformation (see Section 5.4). Unfortunately, even using the correct back-transformation produces error; thus, error bars should be used on the corrected histogram. This is more complex, but can be accomplished using a likelihood-based model; the R package 'disperse' used to generate these error bars is given in Appendix F.

Likelihood intervals of transformation

The transformation described in the main text occurs when data drawn from Euclidean space is placed into bins in a one-dimensional histogram. Let the true probability of observing a data point in bin i be called p_i, where $i \in \{1:k\}$. The value of p_i is unknown, and to be estimated from the data. Let the total number of data points be n, and the observed number of points in bin i be denoted m_i. Assuming independence between data points, the observed data m_1 to m_k are simply draws from the multinomial distribution on p_1 to p_k. As such, each individual value m_i is marginally binomial on p_i, and the maximum likelihood estimate of p_i occurs when $p_i = m_i/n$ for all i. Thus, one way of interpreting a histogram is as a maximum likelihood estimate of the underlying probability mass function, rather than as a descriptive plot of the observed data. The advantage to this perspective is that we can improve upon the maximum likelihood estimate – we can calculate the full likelihood function. The likelihood function for bin i is:

$$L(p_i|m_i) \propto p_i^{m_i}(1-p_i)^{n-m_i}$$

(Equation 5.1)

which is obtained by dropping the constants of proportionality from the binomial function. Then, following the advice of Edwards (1974), we can consider the likely range of values to be anything that lies within two support units (that is, two log-likelihood units) of the maximum likelihood. This range can be easily evaluated in R, and the code is available via the supplementary material. When performing the transformation described in the main text it is a simple case of stretching this function over the new range – leading to inflated maximum likelihood estimates and likelihood intervals near the origin.

The geometrical error in assuming that a histogram represents an angular slice of a two-dimensional dispersal distribution might appear to have little relevance to biology, but it can – and frequently does – lead to erroneous biological conclusions (see Section 5.5). Consequently, many statements made about – for example – natal dispersal of birds, the breeding dispersal of mammals and the seed shadows of plants may well be in error. Of course, my sample of 142 papers (Table 5.1) does not represent a complete survey of all of the available literature, but does suggest that the mistake I describe is prevalent in the biological literature. In fact, it seems that some areas of study are more prone to this error than others. For example, in plant sciences, where models of seed dispersal are more mathematically advanced than models in other fields, many authors are aware of this problem and the models used

deal explicitly with this issue (see for example Clark *et al.* (1999) and Venable & Lawlor (1980)). However, as these models are mathematically complex, many field ecologists may well be applying the results without a full understanding of these processes: in some cases, a complex model is used but is presented with a histogram which has not been correctly transformed (Higgins *et al.* 2003).

Rule 3: Biological inferences drawn from small data sets should be treated with caution

Even when data are correctly back-transformed and appropriate error bars included (see Rules 1 and 2), care should be taken drawing biological inferences from data sets with small sample sizes, since in these cases the size of the resulting error bars is larger and can swamp any apparent patterns. For example, a paper describing dispersal in the great bustard, *Otis tarda*, explicitly states that 'The juvenile dispersal period was longer and the distances reached farther in males than in females. Natal dispersal distances were also longer in males, all of which dispersed from their natal areas and established as adults at 5-65 km from their natal nests. In contrast, most females were strongly philopatric, settling at 0.5-5 km from their natal nests. These marked sex differences in offspring independence and dispersal may have evolved originally to maintain genetic diversity and are probably reinforced through male competition for mates' (Alonso *et al.* 1998). In fact, the small sample sizes (1, 5, 2, 2, 2 and 1 in each bin of the histogram in Figure 3 from this paper) mean that, even when the data are correctly transformed, the resulting error bars completely obscure the pattern from which these conclusions are drawn (Figure 5.4). Similarly, the transformation has the greatest impact close to the source location. If data are biased

towards sites close to the source, there will be more change introduced by the transformation. I strongly caution authors who have small sample sizes to be aware that it is very difficult to make strong claims about the nature of dispersal in such a situation.

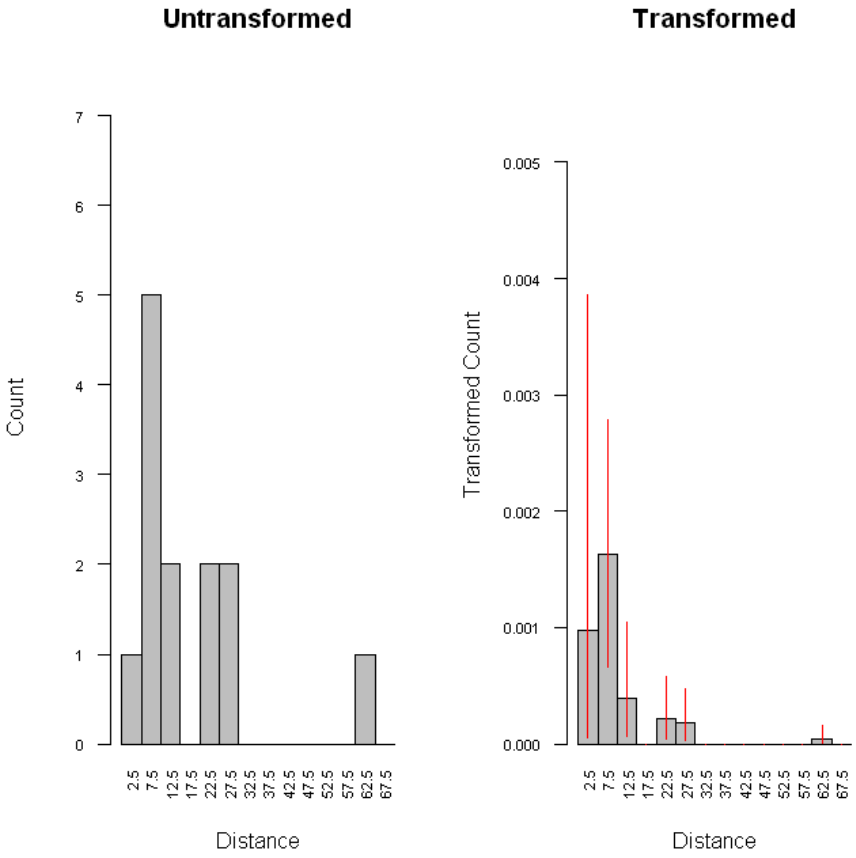


Figure 5.4 (a) Dispersal in male great bustards, *Otis tarda*, redrawn from Figure 3 of Alonso *et al.* (1998), apparently showing that – in contrast to females – males disperse from their natal areas (there is an area of low counts near the natal site). (b) When the same data are correctly transformed and appropriate error bars added, the apparent pattern difference disappears.

5.4 Transforming two-dimensional map data into one-dimensional histograms

For any bivariate probability density function $f_{X:Y}(x;y)$ that has radial symmetry, the density function of the radius $f_R(r)$ where $R = \sqrt{X^2 + Y^2}$ is obtained by the following simple transformation (essentially multiplying through by 2πr as we would expect):

$$f_R(r) = 2\pi r f_{X,Y}(x, \sqrt{r^2 - x^2})$$

(Equation 5.2)

As an example, let X and Y be independent normal random variables, meaning the probability density function $f_{X:Y}(x;y)$ is as follows:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2} exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\}$$

(Equation 5.3)

then by multiplying through by 2πr and substituting $y = \sqrt{r^2 - x^2}$ we arrive at the Rayleigh distribution:

$$f_R(r) = \frac{r}{\sigma^2} exp\left\{-\frac{r^2}{2\sigma^2}\right\}$$

(Equation 5.4)

This holds true for all bivariate distributions with radial symmetry. The next challenge is to recover the shape of the bivariate distribution by doing some sort of back transformation. Note that we can never get back to the original Cartesian coordinates from the distribution of radii alone, as we have lost information when we condensed the x and y coordinates into a single variable r (in order to get back we would also require the distribution of angles).

For any radially symmetric distribution this angular slice will have the same form as $f_{X:Y}(x; 0)$, the only difference being that our slice is defined over (0, ∞) rather than (-∞,∞). Thus, if we denote the distance from the centre along the angular slice as r*then:

$$f_{R^*}(r^*) = \frac{f_R(r)}{2\pi r}$$

(Equation 5.5)

$$f_{R^*}(r^*) = f_{X,Y}(x, 0), \qquad 0 \leq x \leq \infty$$

(Equation 5.6)

Since the distribution of $f_{X:Y}(x; 0)$ is symmetric (this being already specified by the

assumption of radial symmetry) we can obtain the full distribution of X, and by extension the full bivariate distribution.

The transformation needs to be modified to be useful across the space defined by the histogram. In a histogram, let the total number of observations be called *n*, and the total number of bins be called *k*. Let the number of points in bin *i* be called m_j, meaning the sum from *i*=1 to *k* over m_i is equal to *n*. Furthermore, we will say that bin *i* starts at a distance α_i from the origin and ends at a distance β_i from the origin (in order to allow flexibility, as it is possible that not all bins will be of equal width). Finally, the unknown height of ring *i* will be called h_i.

Then, the volume of the *i*th ring is equal to $h_i*\pi*(\beta_i^2 - \alpha_i^2)$. Which should be possible to equate to the observed density of the histogram; in other words to m_i. Therefore, the height of the *i*th ring is given by $h_i = m_i/[\pi(\beta_i^2 - \alpha_i^2)]$. This is the general form of the transformation that we should use. If we plot a histogram of the h_i values then we will be actually reconstructing what we want – something proportional to the Cartesian probability density.

The transformation is perhaps even more intuitive if every bin is of an equal width *s* (in other words $\beta_i - \alpha_i = s$ for all *i*). It follows that $\alpha_i=(i-1)*s$, and $\beta_i=i*s$. The transformation becomes $hi = m_i/[\pi s2(i^2 - (i-1)^2)]$, which can be simplified down to $hi = mi/[\pi s^2(2i^2 - 1)]$. This formulation makes clear the connection to the probability density transformation, which would be to simply divide out the $2\pi r$ that we introduced earlier. As we let the bins become thinner and more numerous the value of *i* increases for the same distance from the origin, and we can see that when *i* becomes large the number in the denominator essentially becomes $2\pi i$ (the s^2 can be thought of as a scaling factor). The R package 'disperse' used to create this

transformation is available in Appendix F.

5.5 Errors in biological interpretation following incorrect transformations

Unfortunately, many authors who present histograms appear to think that they are presenting an angular slice of a dispersal distribution as described in Section 5.4. In fact, failure to correctly transform data from two-dimensional maps to one-dimensional histograms (or vice versa) can lead to incorrect biological inferences.

Here we give just one example from the many we could have chosen. Figure 5.5 shows the dispersal of male and female black kites, redrawn from Figure 3 in Forero *et al.* (2002). A central claim of this paper is that males tend to disperse over short distances, while females disperse over all distance categories equally; the consequences of – and reasons for – these apparent sex differences in dispersal are discussed. In fact, when the data are correctly transformed, the sex differences disappear and it is apparent that both sexes show the same exponential decline in dispersal distance.

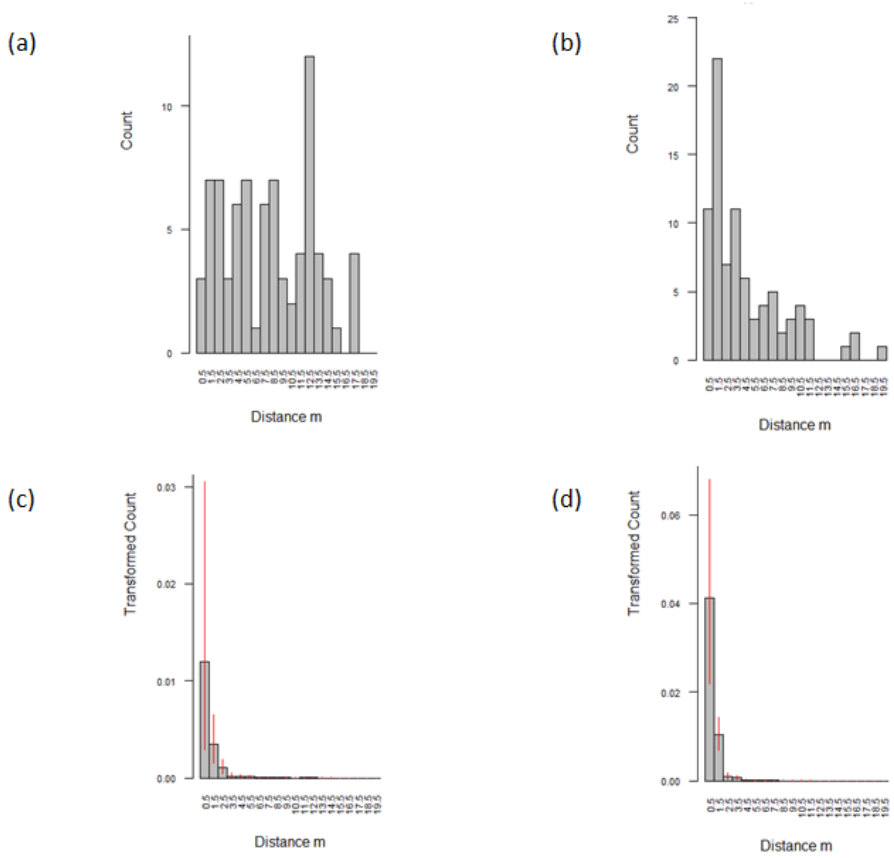


Figure 5.5 (a) and (b) show dispersal in female and male black kites, redrawn from Figure 3 of Forero *et al.* (2002). (c) and (d) show the same data, correctly transformed. The apparent differences (males tending to disperse over short distances, and females dispersing equally at all distance categories) disappear; the two distributions do not differ.

In other cases we see data move from an apparently normal dispersal distribution to a negative exponential; I tested 52 dispersal distributions for normality using a KS test, and in 24 cases the data initially tested were normal prior to transformation but were not afterwards. This is a significant problem as many authors make a range of assumptions about the data based on its normality. For example, many of the results of parametric statistics could be in error, while Bayesian approaches will often fail to produce analytically tractable results outside the bounds of the mathematical ease of

the normal distribution.

5.6 Conclusions

The error I describe is common in the literature, and can lead to major errors in the interpretation of data – which is perhaps especially annoying when the data in question are as difficult and time-consuming to collect as is often the case. We suggest that following the rules we have outlined will first of all help authors present data in the most informative way, and second ensure that biological inferences drawn from these data are accurate.

Chapter 6: Complete applications of GP

6.1 Abstract

I present applications of the complete DPM model presented in Chapter 4 to three separate problems. These problems are found in three separate fields: (i) in ecology, locating the potential sources of invasive newts in the UK; (ii) in epidemiology, with drug-resistant bacteria in East London, and (iii) in criminology, locating the home locations of anti-Nazi dissidents during the Second World War in Berlin. Slight variations to the complete DPM model are applied to each of these cases and these different approaches are summarised and compared.

6.2 Introduction

The previous chapters have shown that the DPM model is an effective tool for assessing the geographic profiling problem. Chapters 3 and 4 demonstrated the model's effectiveness both theoretically and with simulated data. Chapter 5 went on to establish the correct types of dispersal distributions used in real world applications of this model. This chapter brings these together by applying the DPM model to real problems in three separate fields, two biological and one criminological. The first application is in the field of invasive species ecology. The DPM model is used to locate the potential sources of invasive newts in the UK; this is an exploratoory study, with GP being used to drive hypotheis building and further data collection. The second application is in epidemiology, where GP is used to study drug-resistant bacteria in East London. This is a classical GP study applied in a novel setting, with the DPM model used to prioritise a list of GP practices to investigate which ones

may be suspects for over-presscription of antibiotics. The third application is in criminology. The DPM model is used to locate the home locations of anti-Nazi dissidents during the Second World War in Berlin. This study was chosen as it is a multi-site problem with large data volumes, in which the Rossmo model might be expected to perfrom poorly (Stevenson *et al.* 2012). This final study also illustrates that developments in this thesis can be fed back into both criminology and biology. Slight variations to the complete DPM model are applied to each of these cases and these different approaches are summarised and compared.

6.3 Alpine newts

Alpine newts: introduction

The alpine newt *Mesotriton alpestris* (Laurenti 1768) was originally found in Central Europe and Southern Europe (Chinery 1996) but has been introduced into Britain since the 1920s (Beebee & Griffiths 2000). Since their arrival in Newdigate in Surrey they have spread to several other surrounding areas. Beebee & Griffiths (2000) illustrate populations existing in five other locations: South East London, Sunderland, Shropshire, Birmingham and Brighton. Most recently, in 2011, the Non-native Species Secretariat described Alpine newts as being established locally at more than 40 sites in the UK (Wilkinson 2011). Expert opinion asserts that the Alpine newt has become established due in part to deliberate introductions (Wilkinson 2011).

The recently introduced newts have thrived in this country, reaching high population densities in some areas (Beebee & Griffiths 2000). However, unassisted movement

between ponds seems to be very slow (Beebee and Griffiths 2000; Bond & Haycock 2008; Wilkinson 2011). For instance, although Alpine newts were introduced to ponds in Shropshire in 1970-74 a survey some 20 years later found Alpine newts in ponds up to 70m from the introduction site, with little evidence to suggest colonisation further afield (Bell & Bell 1995). Work by Smith and Green (2005) has suggested that the average dispersal distance for the Alpine newt between ponds in somewhere in the region of 500m.

At the moment, it is unclear whether the establishment of Alpine newts in the UK is a significant conservation issue. However, there is the potential for these invasive newts to compete with native species (e.g. at a Sheffield release site Alpine newts are now reported as the most dominant newt species in the pond, coexisting with smooth and great crested newts (Bond & Haycock 2008)). They can reduce biodiversity (Dick *et al.* 2013) and spread pathogens: the amphibian fungal pathogen *Batrachochytrium dendrobatidis* (Bd) which causes the disease Chytridiomycosis was found in 135 Alpine newts at six UK sites in 2008 (Wilkinson 2011). Alpine newts have been placed under Part I of Schedule 9 of section 14 of the Wildlife and Countryside Act 1981, which makes it an offence to deliberately release these animals into the British countryside (Gunner 1984).

This study uses GP to assess the nature of the Alpine newts spread in the UK. GP can be used to assess the likelihood of possible introduction sources and to determine the nature of the spread of Alpine newts. Specifically I assert: (i) GP can be used to locate the potential sources of an amphibious invasive species; (ii) GP can be used to demonstrate that human mediated release is taking place. Crucially, the DPM's posterior estimate of sigma can be compared to data from a previous survey by Smith & Green (2005) that concluded the dispersal distance of newts was on average

500m. The model's performance can be validated by examining the hit scores obtained for, first, a primary school in Lower Wharfedale that is known to have kept and released newts, and second Edinburgh University (where work on alpine newts is carried out).

Alpine newts: methods

Data collection. The locations of Alpine newt sites were taken from a number of sources (Appendix E), including published reports, ecological surveys and local record centres, and from various Amphibian and Reptile groups (ARGs) and ecological consultancies from around the UK; Trent Garner and Gail Austen-Price at the Institute of Zoology, London played a major role in acquiring these data, which was then collated by Ms Robyn Crowther (QMUL) as part of her final-year research project at Queen Mary University of London. Data were transformed from various different formats to decimal latitude and longitude. Site names and addresses were transformed using Google Earth (2012), while OS grid references were converted using the algorithm found in Annex C of Crossley (1999).

Potential newt source locations were taken from field experts and the DPM model introduced in Chapter 3 used to analyse the data.

Suspect sites. Two suspect sites were identified: Burley Oaks Primary School in Lower Wharfedale, which is known to have kept Alpine newts for educational purposes (Bond & Haycock 2008), and Edinburgh University, where the newts are studied.

Alpine newts: results

Smith and Green's (2005) estimate of a 500m dispersal for alpine newts equates to a sigma of approximately 0.0025, and this was used to generate a prior on sigma that peaked at this level (Figure 6.1). Interestingly, the model's posterior estimate of 0.02 was considerably higher, and equates to a typical dispersal distance of around 2.2 km (Figure 6.1).

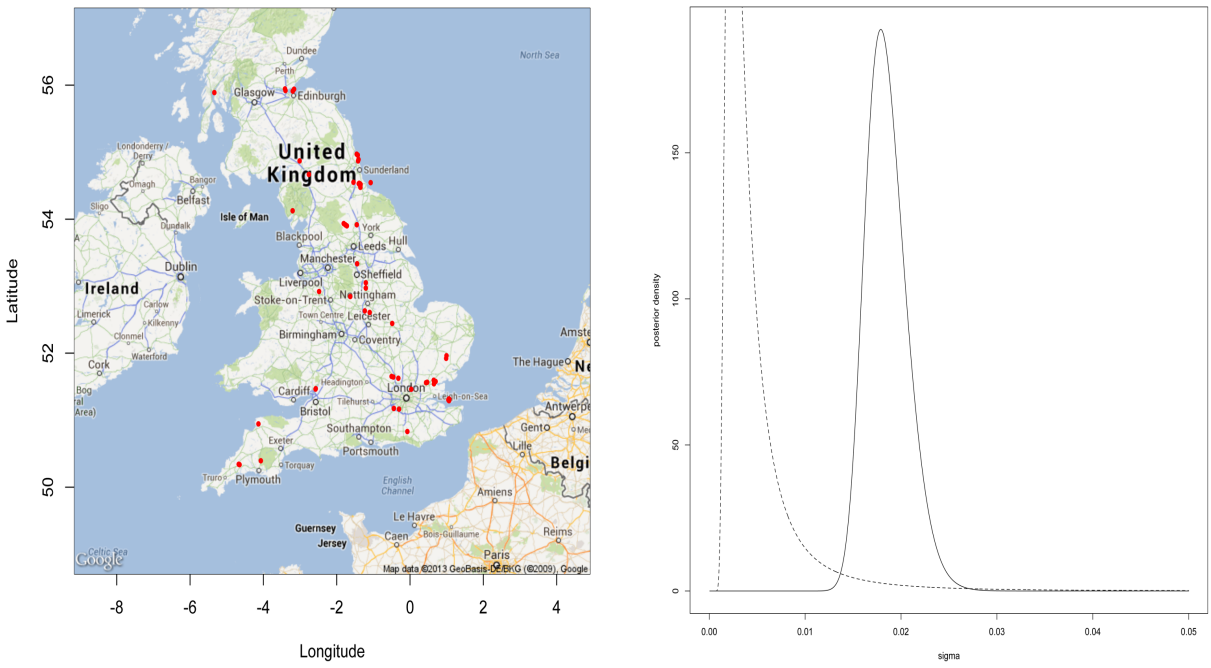


Figure 6.1 Left: The locations of alpine newts across the UK. Right: the prior (dotted line) and posterior estimates of sigma (solid line) for these data.

Across the whole of the UK, the hit scores for the school and university were 0.03% and 0.3% respectively. In fact, at this scale, a large part of the search area is sea, which means that these hit scores are artificially low. However, when the analysis was restricted to just these two areas (with n=6 and n=5 newt locations respectively) the DPM still performed very well, with hit scores of 1.3% and 4.7% respectively

(Figure 6.2).

a) b)

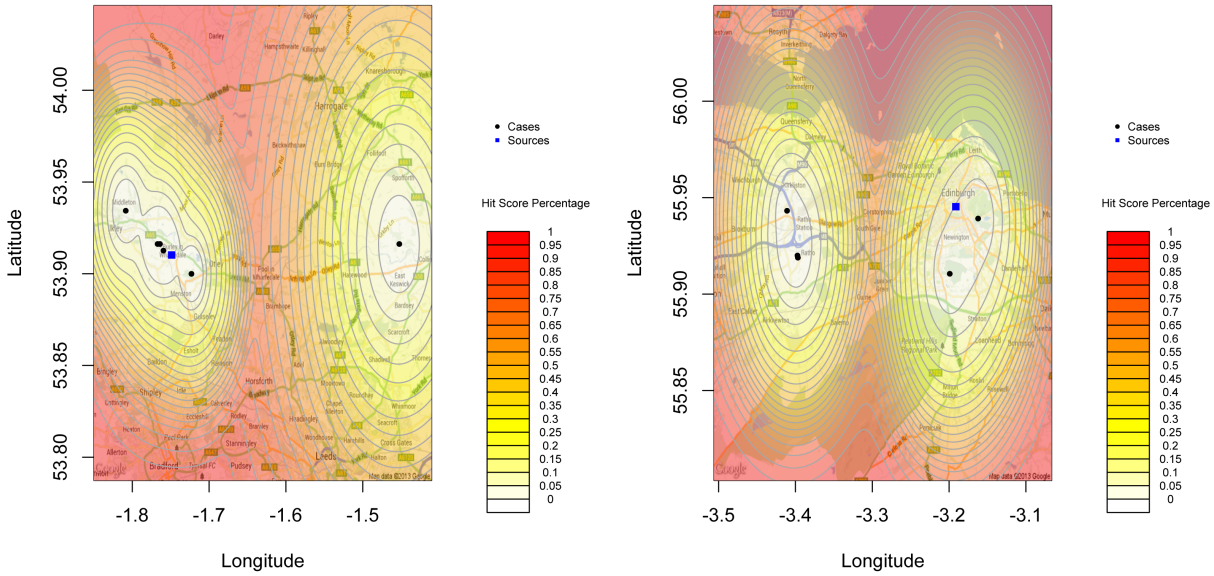


Figure 6.2 Zooms of the central area of the geoprofile produced when the analysis is restricted to (a) the six locations around Burley Oaks Primary School in Lower Wharfedale, which is thought to be the source of these populations, and (b) the five sites around Edinburgh University, another likely source population. In both cases contours show 5% increments. The hit scores for the school and university are 1.3% and 4.7% when the data are restricted to these smaller areas.

Alpine newts: discussion

GP located the potential sources of an aquatic invasive species, bringing search areas down from the entire area of the UK down to tens of kilometres even when the analysis was carried out on small data sets and across relatively small areas. Further improvements such as including only suitable habitat will only increase the effectiveness of the model, allowing precise estimates of potential sources locations

down to the hundreds or even tens of meters.

Interestingly, the model's posterior estimate of sigma suggests dispersal distances much higher than those recorded by Smith & Green (2005), which showed the travel distance of newts to be approximately 500m on average. One solution for the greater dispersal suggested by the DPM – perhaps up to 2.2 km – is that human-mediated movement is involved. Alpine newt movement without the interaction of humans is, as Smith & Green (2005) suggest, probably of short distance. Anecdotal evidence from researchers in the field confirms this (Garner, pers. comm.), yet the newts are clearly spreading further and faster than this would suggest. Human-mediated dispersal would increase the mean dispersal distance fitted from the between-cluster distances. The fact that the average movement of these animals could be around four times higher due to interactions with humans is not unheard of, and a few long distance human-mediated dispersal events could dramatically increase the average distance of dispersal. GP provides a strong confirmation that human-mediated movement is taking place in this species

The results of this analysis are only preliminary, and when made available to experts this early analysis will help to frame further questioning. I recommend that amphibian special interest groups should be made aware of the technique and the findings, and that more potential source locations should be explored, with suspect sites followed up with further investigation. Future work aimed at studying newt dispersal should also consider the importance of human-mediated dispersal, as suggested by the DPM model. In addition, I note that this data set would be well suited to some of the improvements suggested in Chapter 7, in particular habitat-based priors and interactions with SDMs.

6.4 Drug-resistance in the East End of London

Drug-resistance in the East End of London: introduction

Antimicrobial resistance is considered one of the major public health concerns and is responsible for 25,000 deaths in Europe annually and costs the European economy in excess of €1.5 billion a year (Schito *et al.* 2000). It is now estimated that in certain types of infections resistance to common antibiotics may be as high as 70% (Klugman 1990). Gram negative bacteria have been shown to have an increased ability to become resistant to many antibiotics (common organisms include *Klebsiella pneumoniae* and *Escherichia coli)*, and certain strains of gram negative *Escherichia coli* are resistant to more than 90% of naturally derived antibiotics (Nikaido 1998). This issue has expanded beyond hospital-acquired infections and is now increasingly prevalent in the community (Paterson 2006): a surveillance study conducted by Calgary Health Region of Canada found that 71% of reported extended spectrum beta lactamases (ESBL) *E. coli* infections had originated in the community (Paterson 2006). GP is a tool that would enable the identification of potential hotspots responsible for the spread in the community, would allow these hotspots to be identified and eradicated. This could prove to be essential in preventing deaths as a result of community-acquired infections.

In this study I will use GP as a spatial tool to identify potential hot spots responsible for the spread of community- and hospital-acquired antibiotic resistant gram negative bacterial infections in East London. I will analyse data of patients presenting with community- and hospital-acquired antibiotic resistant gram negative bacteria to the Royal London Hospital in Whitechapel. I focus on extended spectrum beta lactamase (ESBL) producing gram negative bacteria, in particular *Escherichia coli*.

I ask the following questions:

i. Can GP correctly locate the hospital that is the known source of infection from patients with hospital-acquired infections?
ii. Can GP prioritise health centres from patients with community-acquired infections?
iii. Is it necessary to split data as in (i) and (ii) above, or can GP locate both hospitals and health centres from the combined data?

Drug-resistance in the East End of London: methods

Data. Data were acquired from the Department of Medical Microbiology at the Royal London hospital in Whitechapel. The records are from 357 anonymised patients that presented to the Royal London with either community- (296 patients) or hospital-acquired (61 patients) strains of extended spectrum beta lactamase producing gram negative bacterial urinary tract infections. Hospital-acquired (nosocomial) infection is defined as an infection that occur afters the first 48 hours of a hospital visit; thus, it is not an infection that the patient had been carrying or suffering from before coming to the hospital (Duchel *et al.* 2002). Usually the site of infection will also be a determining factor (e.g surgical wound or urinary tract) (Duchel *et al.* 2002). A community-acquired infection is defined as an infection that occurred before the first 48 hours of hospital visit and usually occurred in the community setting, without having had any medical procedures and treatments (Yardena *et al.* 2002). The presence of these infections were confirmed by analysing urine specimens at the Royal London.

Patients with hospital-acquired infections had details of the wards they had been admitted to recorded and patients with community-acquired infection had details of the health centre they visited recorded. The postcodes of these health centres, along with that of the Royal London and other local hospitals, were used as potential suspect sites for the infections. The input data for the geographic profile in this case were the postcodes of the patient's permanent residences. These residences were converted in to latitude and longitude values using the Ordnance Survey's algorithm 3 (Ordnance Survey 2010).

Three data sets were analysed, to address the three questions above: (i) hospital-acquired infections; (ii) community-acquired infections; (iii) all infections.

Model. The DPM algorithm presented in Chapter 5 was applied to the model. In each case five chains were run for 10000 iterations, as significant autocorrelation was found in the posterior draws the results thinned by a factor of 500 before assembling the geoprofile.

Drug-resistance in the East End of London: results

Figure 6.3 shows the results of the GP analysis. In each analysis, searching the top 10% of the geoprofile located the sources of 50-100% of the infections (Table 6.1).

In the analysis of hospital-acquired infections, the hit score for the Royal London was 8.5%, with the main peaks of the geoprofile identifying the Royal London and the Mile End Hospital; a number of health centres – among them XX Place Surgery, which is known to be the source of a large number of community-acquired infections (Table 6.2) – also feature in this peak. Indeed, the complexity of the surface suggests

that the pattern of infection may be more complex than simple hospital infection.

The community-acquired infections showed a complex pattern (Figure 6.3). The model performed particularly well in picking out the Royal London, Island Health and St Pauls Way Medical Trust in these five sites, but other important suspect sites – such as XX Place and the Limehouse Practice – fell much further down the geoprofile. However, the model located the sources of more than 50% of infections in the top 10% of the geoprofile and – perhaps more impressively, given the large target area – the top five suspect sites (out of 38) accounted for 115 out of the 296 infections.

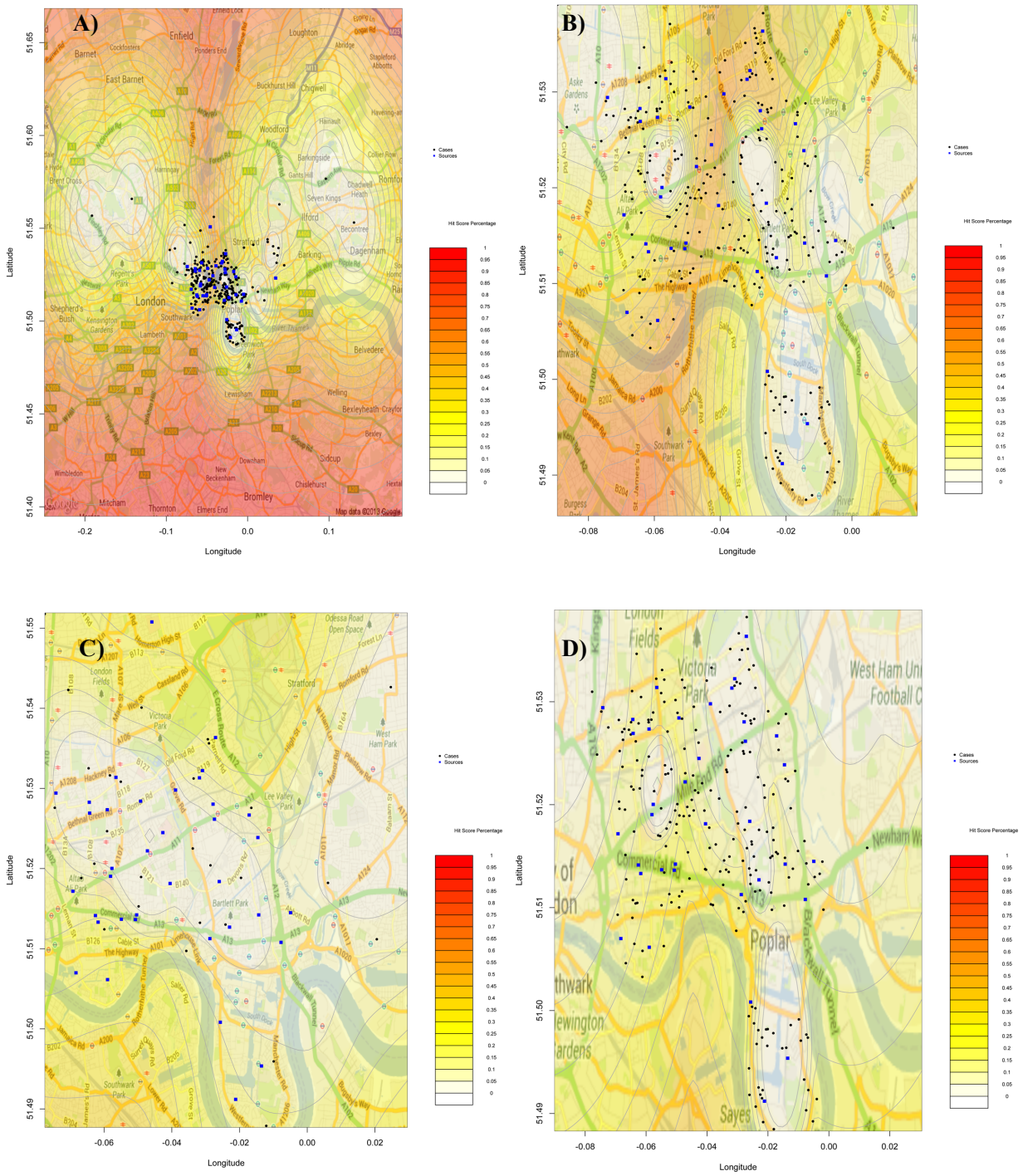


Figure 6.3 Disease locations, suspect sites and jeopardy surface for the study of ESBL antibiotic resistance in the East End of London. (A) The full geoprofile, using all infections. (B) Hospital-acquired infections only, in zoom view. (C) Community-acquired infections only, again in zoom view. (D) All infections, in zoom view. In all cases, contours show 5% increments. Patient addresses are shown as black circles and health centres/hospitals as blue squares.

Table 6.1 Total cases of ESBL resistance, showing the number of infections whose sources occur in the top 1%, 5% and 10% of the relevant geoprofiles, for hospital-acquired infections, community-acquired infections and for all infections combined.

Data	Total cases	Cases in top 1% of surface (%)	Cases in top 5% of surface (%)	Cases in top 10% of surface (%)
Hospital-acquired infections	61	0	0	61 (100%)
Community-acquired infections	296	46 (15.5%)	115 (38.9%)	153 (51.7)
All	357	57 (16.0%)	188 (52.7%)	189 (52.9%)

The analysis of all infections combined was very similar to the analysis of the community-acquired infections, in part because most infections were of this type. Here, the top eight sites in the analysis accounted for 188 out of 357 infections, all in the top 5% of the surface.

Table 6.2 Total cases of ESBL resistance, showing hit scores and numbers of community-acquired, hospital-acquired and total infections for all of the hospitals and health centres in the analysis.

			Hit scores			Number of cases		
Name	Lon	Lat	Hospital-acquired	Community-acquired	All cases	Hospital-acquired	Community-acquired	All
Albion Health Centre	-0.05767	51.52003	0.042	0.000	0.001		7	7
All Saints Medical Centre	-0.00774	51.51079	0.058	0.115	0.047		1	1
Barkantine Practice	-0.02568	51.50082	0.141	0.233	0.105		14	14
Bethnal Green Health Centre	-0.06427	51.52692	0.003	0.119	0.121		11	11
Brayford Square Surgery	-0.05038	51.51422	0.092	0.251	0.496		3	3
Chrisp Street Health Centre	-0.01435	51.51422	0.033	0.116	0.132		10	10
City Wellbeing Practice	-0.06176	51.51331	0.208	0.173	0.337		3	3
Docklands Medical Centre	-0.02114	51.49120	0.173	0.056	0.062		1	1

East One Health	-0.05490	51.51368	0.162	0.222	0.424		5	5
Globe Town Surgery	-0.04923	51.52842	0.001	0.172	0.369		6	6
Gough Walk	-0.02295	51.51270	0.012	0.000	0.002		6	6
Grove Road Surgery	-0.03889	51.52980	0.019	0.035	0.576		2	2
Harley Grove Medical Centre	-0.02782	51.52804	0.083	0.016	0.464		5	5
Homerton Hospital	-0.04590	51.55078	0.262	0.247	0.555			
Island Health	-0.01353	51.49537	0.119	0.001	0.004		19	19
Jubilee Street Practice	-0.05074	51.51363	0.092	0.255	0.471		7	7
Limehouse Practice (Gill St HC)	-0.02876	51.51126	0.039	0.196	0.235		17	17
Merchant Street Practice	-0.02741	51.52616	0.045	0.024	0.283		1	1
Mile End Hospital	-0.04265	51.52449	0.001	0.098	0.536			
Mission Medical Practice	-0.05649	51.53138	0.006	0.161	0.216		6	6
Newham Hospital	0.03449	51.52319	0.015	0.100	0.219			
Pollard Row Surgery	-0.06437	51.52827	0.002	0.117	0.109		3	3
Royal London	-0.05809	51.51904	0.085	0.007	0.049	61	69	130
Ruston Street Practice	-0.02710	51.53635	0.215	0.025	0.475		3	3
Spitalfields Practice	-0.06922	51.51717	0.156	0.118	0.193		5	5
St Andrews Health Centre	-0.01458	51.52387	0.089	0.089	0.168		2	2
St Katherine's Dock Practice	-0.06830	51.50699	0.242	0.173	0.384		2	2
St Pauls Way Medical Centre	-0.02599	51.51838	0.004	0.000	0.001		14	14
St Stephens Health Centre	-0.03092	51.53223	0.156	0.013	0.512		7	7
Stepney Health Centre	-0.04050	51.51814	0.008	0.130	0.498		9	9
Stroudley Walk Health Centre	-0.01715	51.52668	0.115	0.085	0.288		1	1
Strouts Place Medical Centre	-0.07423	51.52942	0.044	0.051	0.008		2	2
The Aberfeldy Practice	-0.00494	51.51449	0.062	0.093	0.005		9	9
The Blithehale Medical Centre	-0.05905	51.52734	0.002	0.148	0.202		8	8
The Tredegar Practice	-0.03184	51.53132	0.131	0.012	0.526		5	5

Wapping Health Centre	-0.05897	51.50615	0.266	0.228	0.485		10	10
Whitechapel Health (Shah Jalal MC)	-0.06256	51.51414	0.197	0.169	0.305		4	4
XX Place	-0.04720	51.52221	0.000	0.187	0.464		19	19
					Total cases	**61**	**296**	**357**

Drug-resistance in the East End of London: discussion

GP was broadly successful in locating the sources of outbreaks of drug resistant bacteria solely from the postcodes of the infected patients, reducing the search area by a factor of between five- and 10-fold that required in a random search. While this is not of huge significance in terms of helping the targeting of disease control, as we already know the location of hospital acquired infections, it does help to further illustrate the effectiveness of GP in locating the sources of data from point patterns, and might perhaps offer the possibility of rapid response to outbreaks of community-acquired infection without the necessity for laboratory analysis of urine or blood samples.

Interesting, the DPM model was able to find the relevant suspect sites even when hospital- and community-acquired infections were analysed together. This too could represent a significant increase in efficiency of locating sources of infection.

This represents a good first test for the application of GP to non-vector borne diseases and complements the work on malaria presented in Chapter 3. This shows that GP can be used to prioritise health centres and that it correctly locates the sources of transmissible bacterial diseases solely from the home addresses of infected individuals.

6.5 Geographic profiling in Nazi Berlin: fact and fiction

Geographic profiling in Nazi Berlin: introduction

As discussed in Chapter 1, cases of serial crime such as murder, rape and arson frequently generate too many, rather than too few, suspects (e.g. 268,000 names in the Yorkshire Ripper investigation in the UK during the late 1970s) (Doney 1990)); hence, suspect prioritisation is critical for major police investigations. The same is true of counterterrorism investigations; in March 2009 the US government's terrorist watch list reached one million names, representing approximately 400,000 unique individuals (Rossmo & Harries 2011). GP has been used since its creation in the 1990s as a method for prioritising such large lists of suspects.

The Rossmo distribution (Figure 1.3) includes two key components: a buffer zone and distance decay. Although the idea of distance decay follows from the nearness principle (Zipf 1950; Rossmo & Harries 2011), its combination with the concept of the buffer zone did not take place until the development of the CGT (criminal geographic targeting) algorithm in the 1990s (Rossmo 2000). It is therefore surprising to find both of these ideas described in Joseph Ditzen's 1947 novel (written under the pen-name Hans Fallada) 'Jeder stirbt für sich allein', published in English as 'Alone in Berlin' (Fallada 2010):

> *The dust-coloured man had pulled out a streetmap of Berlin and pinned it on the wall. Now he stuck in a red flag, exactly over the office block in the Neue Königstrasse. 'You see, this is all I can do for the moment. But over the next few weeks, more and more flags will go up, and where the density is greatest,*

that's where our hobgoblin will be found. Because over time he will wear out, and he won't want to go all that way to drop one of his postcards.'
(Fallada 2010)

... the inspector led the gentlemen back to the map, and, speaking in a whisper, showed them how although there were flags evenly sowed all over the area north of the Alex [the Alexanderplatz], one little area had none at all.
(Fallada 2010)

'And that's where my Hobgoblin lives. He doesn't drop any cards there, because he is too well known; he would have to worry that a neighbour might see and identify him. It's a little working-class enclave, just a couple of streets. That's where he lives.'
(Fallada 2010)

Fallada's novel, which Primo Levi called 'the greatest book ever written about German resistance to the Nazis' (Fallada 2010), is based on the case of Otto and Elise Hampel. After Elise Hampel's brother was killed in France, the Hampels began leaving postcards denouncing the Nazis in apartment buildings around Berlin. The Hampels were arrested in October 1942, tried, and then executed in Plötzensee Prison in 1943.

In fact, both ideas – distance decay and the buffer zone – were used in the Gestapo investigation led by a Kriminalsekretär Püschel:

Hauptverbreitungsgebiet ist nach wie vor die Gegend des Wedding, vor allem

die Strassenzüge beiderseits der Müllerstr. Die Fundorte der Hetzschriften lassen nach wie vor nur den Schluss zu, dass der Hersteller bezw. der Verbreiter nur in der Gegend der Müllerstr, etwa in Höhe der Brüsseler und Amsterdamer Str. wohnen kann. (The main focus of distribution remains the area around Wedding, particularly the streets on both sides of Müller Strasse. These sites at which the inciteful writings were found still suggest that the author or distributor must live in the vicinity of Müller Strasse, probably between Brüsseler and Amsterdamer Strasse.)
(Stapo IV A 1 C, dated 25.9.42).

After Otto Hampel had been identified as a suspect, Püschel noted the existence of a 'buffer zone' around his apartment:

Die Überprüfung der Vorgänge in Bezug auf die Fundorte und die Person Hampel ergab, dass in Wohngrundstück des Hampel derartige Karten nicht gefunden worden sind. Dagegen sind früher einmal die nächsten beiden Eckgrundstücke Thriner Str. 46 und 48 mit derartigen Hetzschriften belegt worden. (Further enquiry into possible connections between retrieval sites and Hampel revealed that no such cards were found on the premises he is living on. However, cards have been retrieved from neighbouring corner properties 46 and 48 Thriner Strasse [presumably a typographical error for Turiner Strasse].)
(Stapo IV A 1 c, dated 26.9.42)

In this study, I digitised and geocoded the locations of 205 of the 214 individual addresses at which postcards or letters were dropped between 2 September 1940 and 16 September 1942 (Figure 1). Addresses were obtained from the Gestapo file held in the Bundesarchiv in Berlin (file NY-36). The file subdivides these locations into seven Bands, or volumes, based on their temporal order, and geoprofiles were prepared for each volume separately, as well as for all incidents combined (the base case); another analysis, in which duplicate addresses were removed, was also carried out.

Geographic profiling in Nazi Berlin: methods

Crime sites. I took the list of addresses at which postcards and letters were found from the original Gestapo files in the Bundesarchiv, Berlin (file NY-36, 1-4), and used historical maps from http://www.alt-berlin.info/ to identify their exact locations on a modern map of Berlin. Incidents that could not be associated with precise locations (e.g. incident no. 181 is assigned in the Gestapo file to 'Wedding') were excluded.

Suspect sites. Known suspect sites are listed in Table 6.3. Suspect sites were identified from the Gestapo archive, with the exception of the underground stations Schönhauser Allee and Hallesches Tor, which were identified from modern maps of Berlin after the analysis.

Geoprofiling. I analysed the addresses discussed in this article using the GP algorithm as described in Chapter 5. 10000 iterations of the model were used, with a thinning factor of 500.

Table 6.3 Locations and hit score percentages for Otto and Elise Hampel's apartment, with those of relatives and two underground stations thought to be associated with travel to relatives' homes. Alfred Lemme moved house during the investigation, and his two addresses are designated (1) and (2) accordingly. Note that the DPM model is able to distinguish between the Hampels' apartment, and the addresses of Otto Hampel's parents and sister, immediately to the south-west and north-east.

Location Name	Lattitude	Longitude	Hitscores
Otto and Elise Hampel	13.35438	52.549623	0.011
Alfred Lemme (Elise Hampel's brother) (1)	13.440645	52.497	0.019
Alfred Lemme (Elise Hampel's brother) (2)	13.357756	52.498917	0.179
Gustave and Pauline Hampel (Otto Hampel's parents)	13.350156	52.547874	0.102
Anna Bartnick (Otto Hampel's sister)	13.357407	52.552048	0.179
Franz Honisch (Elise Hampel's brother-in-law)	13.225681	52.622505	0.739

Geographic profiling in Nazi Berlin: results

The suspect sites and hit scores for the DPM model are shown in Table 6.3. The Hampels' apartment, at Berlin's district N 65, Amsterdamer Straße 10, occurs within the top 1.4% of the geoprofile (when all cases are used), and in the top 6% of the geoprofile for all seven volumes; for four of these, it lies in the top 1% of the surface. Strikingly, analysing the first band, with just 34 locations, shows that the data were sufficient to locate the Hampels' apartment with a high degree of accuracy (0.02%) as early as March 1941, reducing the hunting area to 0.01 square miles – an area of roughly 100m by 100m (Table 6.4; Fig. 6.5). Despite the Gestapo's reputation for efficiency (Gellately 1992), modern GP methods – and the DPM in particular – are thus a considerable improvement on the original investigation, in

which the Hampels were only arrested after two years and 214 incidents.

Table 6.4 Hit scores, incidents, hunting area and target area for each of the scenarios investigated using the DPM. *The hit score describes the percentage of the hunting area (the area encompassing all crime sites, plus a 'guard rail' of 5%) that must be searched before the correct source (in this case, the Hampels' apartment) is located. †'Incidents' shows the number of incidents analysed in each band, with the total number of incidents in that band in brackets. Some incidents were excluded because the Gestapo file contained insufficient information to locate them. § The target area is given by the hit score multiplied by the hunting area.

Scenario	**Hit score***	**Incidents†**	**Hunting area (square miles)**	**Target area (square miles) §**
Volume I (2 September 1940 to 11 March 1941)	0.0002	34 (35)	34.14	0.01
Volume II (12 March 1941 to 6 April 1941)	0.0002	33 (34)	82.60	0.02
Volume III (12 April 1941 to 5 June 1941)	0.009	30 (32)	144.60	1.30
Volume IV (4 June 1941 to 24 August 1941)	0.054	31 (32)	13.22	0.71
Volume V (31 August 1941 to 28 December 1941)	0.001	34 (35)	24.37	0.02
Volume VI (1 February 1942 to 30 May 1942)	0.014	33 (35)	43.78	0.61
Volume VII (12 July 1941 to 16 September 1942)	0.064	10 (12)	54.18	3.47
Base case (all sites)	0.014	205 (215)	146.44	2.05
No duplicates	0.009	190	146.78	1.32

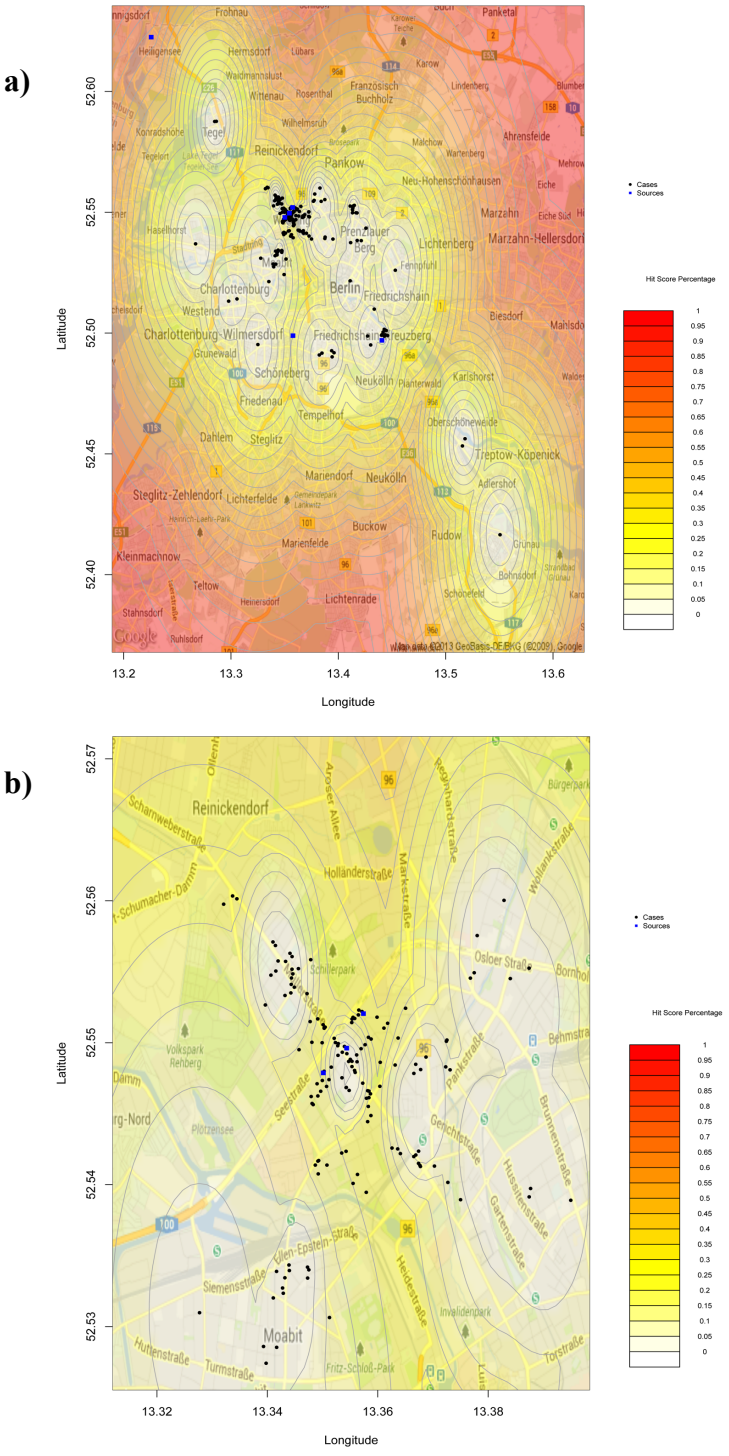


Figure 6.4 Crime sites (black dots), suspect sites (blue squares) and jeopardy surface for the Hampel case. Black circles show the locations at which postcards and letters were dropped by the Hampels between September 1940 and September 1942. Blue squares mark 'suspect sites' – the Hampels' apartment, the addresses of relatives and underground stations. (a) The full geoprofile for the DPM model. (b) Close-up of the peak of the geoprofile. All contours show 5% increments.

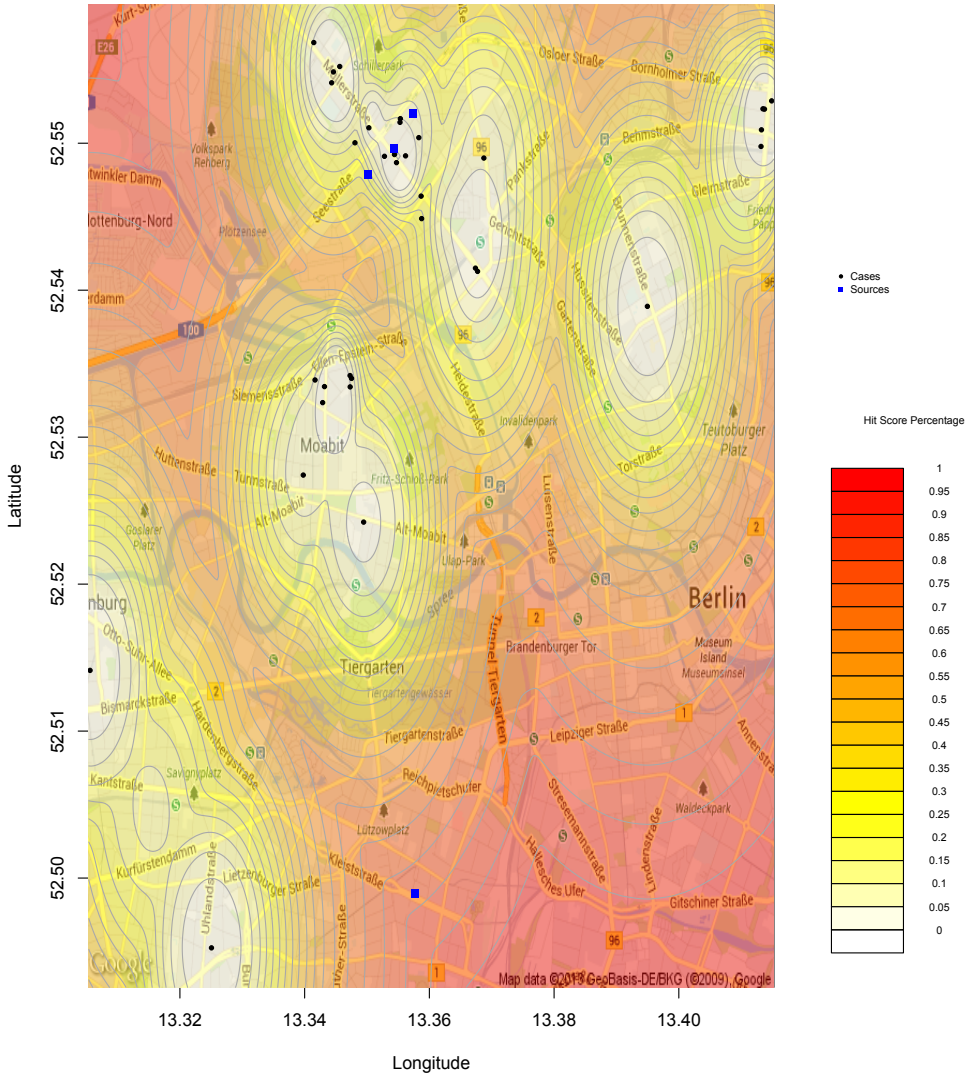


Figure 6.5 The geoprofile obtained by analysing the 34 locations in Band I, showing that the model could have identified the Hampels' home (blue square), with a very high degree of accuracy (0. 02%), as early as March 1941. Contours show 5% increments.

The addresses of other family members also had high hit score percentages; the original home of Alfred Lemme (Elise's brother; 1.9%) in Falkensteinstraße fell in a secondary peak south-east of the main peak. Gustave and Pauline Hampel (Otto's parents; 10.2%) lived very close close to the Hampels, as did Otto's sister, Anna

Bartnick (17.9%), but the accuracy of the DPM is sufficient to differentiate these from the Hampels' home. Secondary peaks around the underground stations at Schönhauser Allee and Hallesches Tor seem to indicate plausible travel routes between the Hampels' apartment and Alfred Lemme's original home, and his second home: travelling from Wedding, the nearest station to the Hampels' apartment, to Schlesisches Tor, near Alfred Lemme's home, the most likely route involves changing underground lines at Schönhauser Allee, while the route from Wedding to Bülowstrasse, near Alfred's home during the later part of the investigation, requires a change at Hallesches Tor.

Geographic profiling in Nazi Berlin: discussion

Beyond its historical interest, our new analysis of the Hampel case demonstrates the potential of GP in similar situations today. The problems that faced the Gestapo have parallels in modern counterterrorism and counterinsurgency efforts, which also have to handle problems of information overload (Rossmo & Harries 2011). This is exactly the problem GP is designed to address by prioritising such large lists of suspects in a meaningful way. In fact, while attention is typically focused on their major attacks – bombings, kidnappings, hijackings – terrorists often engage in low level activities similar to the Hampels' campaign, including vandalism, graffiti, and anti-government leaflet distribution and banner posting (Jordan & Horsburgh 2005). Rossmo and Harries (2011) suggested that the creation of geospatial databases of terrorism-related graffiti could be used to help locate terrorist bases before more serious incidents occur, and our study provides fascinating empirical support for this idea – even if, in the case I describe here, my sympathies are with the insurgents.

Chapter 7: Conclusions

7.1 Abstract

GP continues to change and develop to meet new challenges. Here I outline the current state of GP in biology, given the work described in the previous chapters. I go on to outline the possible future developments to the GP approach and finish by outlining the possible use of GP outside of academia.

7.2 GP in biology

GP has served as useful practical tool in criminology and military science. The ability to locate the home locations of serial offenders from a range of different areas has been invaluable to many existing and ongoing police and military operations, but moving GP from criminology into biology required a thorough examination of the model. In Chapter 1 I address weaknesses of the standard criminological model and expanded on the strengths of this model.

In Chapter 2 I consider the mathematical problems of the current model by discussing model fitting, and then test this new approach with real and simulated data. I show that it is possible to arrive at a more rigorous likelihood-based approach to fitting the GP model, and that this outperforms existing approaches such as kernel density estimate. This explicitly addresses weaknesses (ii) and (iii) outlined in Section 1.7. I go on to test GP on a dataset of 53 invasive species, where it again outperforms simple spatial methods. The work described in this chapter was the first major study of the GP method across a range of datasets in a biological application

(Stevenson *et al.* 2012).

Chapter 3 continues the work started Chapter 2, specifically the examination of the mathematical and theoretical underpinning of GP. I continued the work of O'Leary, expanding the Bayesian GP approach to explicitly deal with multiple sources and unifying the work of O'Leary and Rossmo under one paradigm. I test this new DPM model extensively in simulations and on a real-world malaria dataset. GP again shows that it is effective in a range of settings, confirming strengths (ii) and (iv) in Section 1.7. Epidemiological problems represent a huge growth area for GP where the model could be used to prioritise disease control and potentially save many hundreds of thousands of lives.

Chapter 4 uses further simulations to compare GP to other models, and looks at methods of fitting the parameter value of sigma from the data alone. This builds on the existing strengths (i) and (iii) in Section 1.7. I show that GP performs well under a range of possible simulated settings even when the type of distribution used did not match that from which the data were drawn, confirming the results of O'Leary (2012b).

In Chapter 5 I derive functional dispersal profiles from ecological data, allowing GP to be applied easily to biological problems. In so doing, I reveal a common geometric mistake made by many authors that study or use dispersal information and show that the vast majority of biological data analysed conforms to either a normal distribution or an exponential decay function, the latter supporting the work of Nekola & White (1999) and Nathan & Muller-Landau (2000). I recommend that GP models in biology are fitted using either a normal or an exponential distribution.

Finally, I use the fully formed GP methods in three widely different examples from

three separate fields. These cases serve as outstanding examples of the flexibility of the method as they are drawn from the fields of criminology/history, invasion ecology and epidemiology. GP performs well in all cases, correctly locating the sources of Nazi resistance, invasive newt introduction sites and drug-resistant bacteria outbreaks. Applying GP to more biological settings addresses weakness (i), as there is now an established body of work using GP within biology.

7.3 Future developments

This thesis has taken GP from a relatively obscure technique used little in biology, with little mathematical foundation and limited testing, to become a well-developed method used by other authors, tested extensively with simulation and built on a solid Bayesian framework. However, the work on GP in biology is far from complete. I have identified several key areas of future development and testing:

(i) Genetic data

(ii) SDM/niche models

(iii) Prior information and GIS

(iv) Further applications

(v) Temporal information

(vi) Future prediction

Genetic data

At the time of writing I am working in collaboration with Robert Verity to produce a novel version of the DPM model that will incorporate genetic information. This

model will separately weight the grouping by both spatial information and by genetic data. This mixture clustering approach could also be used to infer the spatial point pattern given genetics or infer the genetics given the spatial point pattern. This model would allow GP to be applied to a whole additional suite of biological problems.

SDM and niche models

The interaction with SDM and niche models still needs developing. SDM models provide an estimate of where a species is likely to be given which habitat they normally occupy. Linking in this information as an informed prior (iii) would equip a GP model with two sets of vital information for locating sources of populations, firstly where the organism in question is likely to live (its preferred habitat from the SDM) and secondly where it is now and how it moves (the GP dispersal data). Together these can produce what several authors have called a next generation SDM model, one that is mathematically robust and built from two models with solid past applications.

Prior information and GIS

Building in GIS data is an important practical consideration for a GP model. Many academics – and even more workers in industry or government – use GIS tools. The various software packages such as ARC GIS (ERSI, 2011) are commonly used worldwide. Allowing GP to interact with these various programs and to use the output they produce would open up the approach to many more people. At the time of writing GP can now incorporate and run shape files using the R packages sp

(Pebesma & Bivand 2005) and maptools (Bivand & Lewin-Koh 2013).

Further applications

GP has now been applied to several areas but there are still many more possibilities. Firstly, I would recommend further work on existing areas: invasive species (especially plant species where the distributions of seed shadows are well studied); epidemiology (where it has already been applied to vector-borne diseases, and could easily be extended to other transmissible diseases which require extended contact such as tuberculosis and legionnaires disease); and animal foraging. The last of these is where GP was first applied in biology (Le Comber *et al.* 2006), and this work could be extended to extremely endangered terrestrial species such as the giant panda and tiger. Furthermore, GP can also be applied to yet more problems such as determining if marine protected areas are the sources of local repopulation or informing ecological age structured population models.

Temporal information

Thus far the only real consideration of temporal trends in geoprofiles was in Chapter 2, where parameter estimates of *B* were compared across different years. There is much more possible work to do in this area, producing geoprofiles that alter in response to new information as time progresses; producing a 4D surface that alters through time would interesting. There is the possibility of some dispersal patterns being correlated in space and time, for example seed emergence varies in both space and time, and seeds dispersed closer to the host plant wait longer to emerge (Venable

& Lawlor 1980), suggesting that this may be an interesting area to explore.

Future prediction

O'Leary (2012) has already made the point that the Bayesian GP could be easily rewritten to run forward in time, predicting the probability of new crimes from the knowledge of existing crimes. This has yet to be tested, and it would be important to test this new approach both with real data and with simulations.

There are many areas of GP still left to develop. I hope that others will continue to work on and improve these approaches as they offer a genuine opportunity to better use limited resources and discover information from point pattern data.

7.4 GP use outside of academia

It is my most sincere hope that a number of bodies and organisations outside of academia take up and use this tool. GP began as an academic endeavour in criminology, but has rapidly moved to become an applied tool used by law enforcement agencies the world over. I hope the same will be true of GP in biology. In particular I highlight three areas that would stand to benefit most from the application of GP:

(i) Public health bodies

(ii) Wildlife management agencies

(iii) Conservation NGOs

These three sectors, while differing in application, often share the same concerns as police forces. Public health bodies must locate the sources of transmissible diseases, including vector-borne diseases such as malaria, dengue fever and sleeping sickness, and GP has already been shown to be effective in doing this. Wildlife management agencies must deal with invasive species emerging from multiple introduction sites, possibly across very large spatial scales. Conservation charities might need to locate the sources of dangerous animals that come into conflict with humans, or determine if poachers are operating within certain protected areas. All of these organisations share the same problem with the police; they have large remits, cover large areas, with a large number of potential suspects, using limited resources. All could benefit from the application of one or another form of GP.

My recommendation is to follow the example in criminology and set up a training program for non-academic experts in these organisations that wish to apply GP to their particular problem. A system of accredited training similar to the Geographic Profiling Crime Analyst (GPA) training program set up by ECRI would provide a pool of skilled experts who could use this method in practical problems that need immediate solutions.

7. 5 Concluding statement

This thesis has developed and expanded GP, taking a practical tool from one field – with some existing mathematical issues – and applying it to interesting and pressing biological problems. I think that this is an exciting and important area of research, with numerous applications to pressing global concerns, from the spread of disease and invasive species to conservation and habitat management. Other authors have

already begun to take up and apply this work, with recent publications on algae invasions in the Mediterranean by Papini *et al.* (2013), as well as ongoing collaborations with Kim Rossmo at Texas State and Trent Garner at the Institute of Zoology. I hope that this momentum continues and GP takes its place within the family of modeling techniques available to ecologists, epidemiologists, wildlife managers and conservation practitioners.

References

A

Alonso, J.C., Martin, E., Alonso, J.A., Morales, M.M. (1998) Proximate and ultimate causes of natal dispersal in the great bustard *Otis tarda*. *Behavioral Ecology,* 3(9): 243-252.

Amsler, C. D. and Searles, R.B. (1980) Vertical distribution of seaweed spores in a water column offshore of North Carolina. *Journal of Phycology* 16: 617-619.

Anderson, E. K. and North, W.J. (1966) In situ studies of spore production and dispersal in the Giant Kelp, *Macrocystis*. Proceedings of the International Seaweed Symposium, 5: 73-86.

Andrew, N. L. and Viejo, R.M. (1998) Ecological limits to the invasion of *Sargassum muticum* in northern Spain. *Aquatic Botany*, 60: 251-263.

Applegate, R. D. (1977) Long-distance homing of a cottontail. *American Midland Naturalist,* 97: 22-1.

Avlles, L and Gelsey, G. (1998) Natal dispersal and demography of a subscial *Anelosimus* species and its implications for the evolution of sociality spiders. *Canadian Journal of Zoology*, 76: 2137-2147.

B

Baker, H. G. (1986) Patterns of plant invasion in North America. In *Ecology of biological invasions of North America and Hawaii* (pp. 44-57). Springer, New York.

Baker H.G. and Stebbins G.L. (1965) *The genetics of Colonizing Species*. Academic Press, New york.

Beebee, T. J. C; and Griffiths, R. A. (2000) *Amphibians and Reptiles in Britain*. Harper Collins, London

Beer, J. R. (1955) Movements of tagged beaver. *Journal of Wildlife Management*, 19, 492-493.

Belisle, M., and St. Clair, C. C. (2001) Cumulative effects of barriers on the movements of forest birds. *Conservation Ecology*; 5(2).

Bell, B. D; and Bell, A. P. (1995) Distribution of the Introduced Alpine Newt (*Triturus alpestris*) and of Native *Trituraus* Species in North Shropshire England. *Australian Journal of Ecology*, 20:367-375.

Belthoff, J. R. and Ritchison G. (1989) Natal Dispersal of Eastern Screech Owls. *The Condor,* 91, 254-265.

Bequaert, J. C. (1943) The genus *Littorina* in the western Atlantic. *Johnsonia,* 1:1-28.

Bivand, R. and Koh N. L. (2013) *maptools: Tools for reading and handling spatial objects*. R package version 0.8-24. <http://CRAN.R-project.org/package=maptools>.

Blackburn, T. M., Pyšek, P., Bacher, S., Carlton, J. T., Duncan, R. P., Jarošík, V., ... & Richardson, D. M. (2011) A proposed unified framework for biological invasions. *Trends in Ecology & Evolution*, 26(7), 333-339.

Blakesley, J.A., Anderson, D.R. and Noon, B.R. (2006) Breeding Dispersal In the California Spotted Owl. *The Condor*, 108: 71-81.

Bloom, T.C., Baskin, J.M. and Baskin, C. C. (2002) Ecological Life history of the facultative woodland biennial *Arabis laevigata variety laevigata* (*Brassicaceae*): Seed Dispersal. *Journal of the Torrey Botanical Soceity*, 129(1): 21-28.

Bond, I., and Haycock, G. (2008) The Alpine newt in northern England. *Herpetological Bulletin*, 104: 4-6.

Bowers, G. L. (1954) An evaluation of cottontail rabbit management in Pennsylvania. *Transactions of the North American Wildlife Conference*, 19:358-367.

Bovet, J. (1980) Homing behaviour and orientation in the redbacked vole, *Clethrionomys gapperi*. *Canadian Journal of Zoology*. 58: 754-760.

Burgess, E. W., & Park, R. E. (1921) *Introduction to the Science of Sociology* (No. s 735). Chicago: University of Chicago Press.

Bray, Y., Devillard, S., Marboutin, E., Mauvy, B. and Pe´roux, R. (2007) Natal dispersal of European hare in France. *Journal of Zoology*, 273: 426–434.

Brantingham, P.J. and P.L. Brantingham (eds.) (1981). *Environmental Criminology.* Beverly Hills, CA: Sage Publications.

Brown, R. O., Rossmo, D. K., Sisak, T., Trahern, R., Jarret, J. and Hanson, J. (2005) *Geographic Profiling Military Capabilities*. Final report submitted to the Topographic Engineering Center, Department of the Army. Fort Belvoir, VA.

Buscema, M., Grossi, E., Breda, M., Jefferson, T. (2009) Outbreaks source: A new mathematical approach to identify their possible location. *Physica A: Statistical*

Mechanics and its Applications, 388:4736–4762.

C

Cain, M. L., Damman, H. and Muir, A. (1998) Seed dispersal and the holocene migration of woodland herbs. *Ecological Monographs*, 68: 325-349.

Canter, D., Coffey, T., Huntley, M., & Missen, C. (2000) Predicting serial killers' home base using a decision support system. *Journal of Quantitative Criminology*, 16: 457–478.

Canter, D., & Hammond, L. (2006) A comparison of the efficacy of different decay functions in geographical profiling for a sample of US serial killers. *Journal of Investigative Psychology and Offender Profiling*, 3(2): 91-103.

Capone, D. L., & Nichols, W. W. (1976) Urban structure and criminal mobility. *American Behavioral Scientist*, 20(2): 199-213.

Carey, J. R. (1996) The incipient Mediterranean fruit fly population in California: implications for invasion biology. *Ecology*, 77(6): 1690-1697.

Carlquist, S. (1965) *Island Life*. Natural History Press, New York

Carlton, J.T. (1985) Transoceanic and interoceanic dispersal of coastal marine organisms: the biology of ballast water. *Oceanography and Marine Biology Annual Review*, 23: 313-371.

Carlton, J. T. and Scanlon, J.A. (1985) Progression and dispersal of an introduced alga: *Codium fragile ssp. tomentosoides* (*Chlorophyta*) on the Atlantic coast of

North America. *Botanica Marina,* 28:155-165.

Carter R, Mendis K, Roberts D. (2000) Spatial targeting of interventions against malaria. *Bulletin of the World Health Organization*, 78:1401–1411.

Caswell, H., Lensink, R., & Neubert, M. G. (2003) Demography and dispersal: Life table response for invasion spread. *Ecology*, 84(8):1968-1978.

Chapman, J. A. (1971) Orientation and homing of the brush rabbit (*Sylvilagus bachmani*). *Journal of Mammalogy*, 52:686-699.

Chew, K. K. (1998) Update on the green crab's movement up the Pacific coast of North America. *Aquaculture Magazine*, July/August: 89-90.

Chinery, M. (1996) *Field Guide to the Wildlife of Britain and Europe*. Larousse, London.

Clark, J. S., Silman, M., Kern, R., Macklin, E., & HilleRisLambers, J. (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology*, 80(5), 1475-1494.

Clements, F.E. (1904) *The development and structure of vegetation*. The Botanical Seminar.

Clobert, J., Baguette, M., Benton, T. G., & Bullock, J. M. (Eds.) (2012) *Dispersal ecology and evolution*. Oxford University Press.

Colautti R. I., and MacIsaac, H. J. (2004) A neutral terminology to define invasive species *Diversity and Distributions*, 10:135-141.

Crawley, M.J. (1987) What makes a community invasible? *Colonization, succession and stability*, A.J. Gray, M.J. Crawley and P.J. Edwards (eds), 429-453. Blackwell,

Oxford, UK.

Crisp, D. J. (1958) The spread of *Elminius modestus* Darwin in north-west Europe. *Journal of the Marine Biological Association of the UK*, 37:483-520.

Cressie, N. A. C. (1991) *Statistics for spatial data*. John Wiley, Chichester.

Crooks, J. & Soulé, M.E. (1999) Lag times in population explosions of invasive species: causes and implications. *Invasive species and biodiversity management* (O.T. Sandlund, S.J. Schei and A. Viken (eds)), pp.103-125. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Crossley, M. (1999) A guide to coordinate systems in Great Britain. *Ordnance Survey. Southampton*
<(http://www.gps.gov.uk/additionalInfo/images/A_guide_to_coord. pdf)>

D

Dale, S, Lunde, A, Steifetten, O. (2004) Longer Breeding Dispersal than natal dispersal in the ortolan bunting. *Behavioural Ecology*, 16(1):20-24.

Danner, D. A., and Fisher A. R. (1977) Evidence of homing by a coyote (*Canis latrans*). *Journal of Mammalogy*, 58:244-245.

Davis, A. R. and Butler, A.J. (1989) Direct observations of larval dispersal in the colonial ascidian *Podoclavella moluccensis Sluiter*: evidence for closed populations. *Journal of Experimental Marine Biology and Ecology*, 127:189-203.

Davies, C. E. *et al.* (2004) *EUNIS habitat classification revised*. European

Environment Agency report
(http://eunis.eea.europa.eu/upload/EUNIS_2004_report.pdf).

Dayton, P. K. (1973) Dispersion, dispersal, and persistence of the annual intertidal alga, *Postelsia palmaeformis* Ruprecht. *Ecology*, 54:433-438.

Deysher, L. and Norton, T.A. (1982) Dispersal and colonization in *Sargassum muticum* (*Yendo*) Fensholt. *Journal of Experimental Marine Biology and Ecology*, 56:179-195.

di Castri, F. (1990) On invading species and invaded ecosystems: the interplay of historical chance and biological necessity. In *Biological invasions in Europe and the Mediterranean Basin* (pp. 3-16). Springer, Netherlands.

Dick, J. T. A; Alexander, M. E; and MacNeil, C.(2013) Natural Born Killers: An Invasive Amphipod is Predatory Throughout its Life History. *Biological Invasions*, 15:309-313.

Diefenbach, D. R., Long, E. S., Rosenberry, C. S., Wallingford, B. D., & Smith, D. R. (2008) Modelling distribution of dispersal distances in male white-tailed deer. *The Journal of Wildlife Management*, 72(6): 1296-1303.

Diggle, P. J. (1985) A kernel method for smoothing point process data. *Applied Statistics,* 34, 138–47.

Doney, R.H. (1990). The Aftermath of the Yorkshire Ripper: The Response of the United Kingdom Police Service. In: S.A. Egger (ed.), *Serial Murder: An Elusive Phenomenon.* New York, NY: Praeger.

Dorazio, R.M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H.L., Jordan, F. (2008)

Modelling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, 64: 635-644.

Douglas, M. J. W. (1970) Movements of hares (*Lepus europaeus*) in high country in New Zealand. *New Zealand Journal of Science*, 13: 287–305.

Drake, J.M and Lodge, D.M. (2004) Effects of environmental variation on extinction and establishment. *Ecology Letters*, 7:26-30.

Drake J.A., Mooney H.A., DiCastri H.A., Groves H. A., Kruger F. J. eds. (1989) *Biological Invasion: a Global Perspective*. New York:Wiley & Sons. 525 pp.

Dramstad, W. (1996) Do bumblebees (Hymenoptera: Apidae) really forage close to their nests? *Journal of Insect Behaviour*, 9: 163-182.

Ducel, G., Fabry, J., & Nicolle, L. (2002) Prevention of hospital acquired infections: a practical guide. *Prevention of hospital acquired infections: a practical guide* (Second edition).

E

Edwards, A.W.F. (1972) *Likelihood*. Cambridge University Press, Cambridge (expanded edition, 1992, Johns Hopkins University Press, Baltimore).

El Said, S., Beier, J. C., Kenway, M. A., Morsy, Z. S., Merdan, A. I. (1986) *Anopheles* population dynamics in two malaria endemic villages in Faiyum governorate, Egypt. *Journal of the American Mosquito Control Association*, 2.

Elith, J., Leatherwick, J. R. (2009) Species distribution models: ecological

explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40: 677-697.

Ellner, S. P., McCauley, E., Kendall, B. E., Briggs, C. J., Hosseini, P. R., Wood, S. N., Janssen, A., Sabelis, Turchin P., Nisbet, R. M. & Murdoch., W. M. (2001) Habitat structure and population persistence in an experimental community. *Nature* 412:538-543.

Elton, C.S. (1958) *The ecology of invasions by animals and plants*. Methuen, London, UK.

Espinoza, J. (1990) The southern limit of *Sargassum muticum* (Yendo) Fensholt (*Phaeophyta, Fucales*) in the Mexican Pacific. *Botanica Marina*, 33:193-196.

ESRI (2011). *ArcGIS Desktop: Release 10*. Environmental Systems Research Institute. Redlands, CA.

F

Fallada H (2010) *Alone in Berlin*. London, Penguin.

Ferguson, T. S. (1983) Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 287-303.

Forero, M.G., Donázar, J.A. and Hiraldo, F. (2002) Causes and Fitness Consequences of Natal Dispersal in a Population of Black Kites. *Ecology* 83, 858-872.

Forsman, E.D., Anthony, R.G., Reid, J.A., Loschl, P.J., Sovern, S.G., Taylor, M.,

Biswell, B.L., Ellingson, A., Meslow, E.C., Miller, G.S., Swindle, K.A., Thrailkill, J.A., Wagner, F.F. and Seaman, D.E. (2002) Natal and Breeding Dispersal of Northern Spotted Owls. *Wildlife Monographs*. 149:1-35.

Friedman, J. and Orshan, G. (1975) *Journal of Ecology*. 63:627-632.

Fritts, S. H.,Paul, W.J. and Mech, L.D. (1984) Movements of translocated wolves in Minnesota. *Journal of Wildlife Management*, 48:709-721.

G

Watson, H. W., & Galton, F. (1875) On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4: 138-144.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996) Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers,* 256-274.

Gellately R (1992) *The Gestapo and German Society: Enforcing Racial Policy 1933-1945* (Oxford University Press, Oxford).

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003) *Bayesian data analysis*, Chapman and Hall/CRC, 2nd Ed.

Gerrodette, T. (1981) Dispersal of the solitary coral *Balanophyllia elegans* by demersal planular larvae. *Ecology*, 62: 611-619.

Giger, R. D. (1973) Movements and homing in Townsend's mole near Tillamook,

Oregon. *Journal of Mammalogy*, 54:648-659.

Goodwin, B.J., McAllister, A.J., and Fahrig, L. (1999) Predicting the invasiveness of plant species based on biological information. *Conservation Biology*, 13:422-426.

Google Inc. (2009). Google Earth (Version 5.1.3533.1731) [Software].

Goeze, E. (1882) *Planzengeographie.* Verlag von Eugen Ulmer, Stuttgart.

Green, P. J., Richardson, S. (2001) Modelling heterogeneity with and without Dirichlet process. *Scandinavian Journal of Statistics*, 28, 355-375.

Gressit J. and Gressit M. (1962). An improved malaise insect trap. *Pacific Insects:* 87.

Grosberg, R. K. (1987) Limited dispersal and proximity-dependent mating success in the colonial ascidian *Botryllus schlosseriI. Evolution*, 41:372-385.

Groves, R. H, di Castri, F. (1991) *Biogeography of Mediterranean invasions* Cambridge University Press.

Gu, W., Novak, R. J. (2009) Agent-based modelling of mosquito foraging behaviour for malaria control. *Transactions of the Royal Society of Tropical Medicine and Hygiene*,103:1105-1112.

Gu, W., Regens, J., Beier, J., Novak R. (2006) Source reduction of mosquito larval habitats has unexpected consequences on malaria transmission. *Proceedings of the National Academy of Sciences*, 103(46), 17560-17563.

Gunner, D. (1984) Wildlife and Countryside Act. *Oryx*, 18(02): 114-114.

H

Haccau, P. Jagers, P., Vatutin, V.A. (2005) *Branching Processes: variation, growth and extinction of populations*. Cambridge University press, Cambridge.

Hamilton, W. J. Jr. (1939) Observations on the life history of the red squirrel in New York. *American Midland Naturalist*, 22:732-745.

Hanski, I. K., Selonen, V. (2009) Female-biased natal dispersal in the Siberian flying squirrel. *Behavioural Ecology*, 20:60-67.

Harrow, Heston. and M. Felson (eds.) (1993).*Routine Activity and Rational Choice.* New Brunswick, NJ: Transaction Books.

Harries, K. (1990) *Geographic factors in policing.* Washington, DC: Police Executive Research Forum.

Harrison, P. G. and Bigley, R.E. (1982) The recent introduction of the seagrass *Zostera japonica* Aschers. and Graebn. to the Pacific coast of North America. *Canadian Journal of Fisheries and Aquatic Science*, 39:1542-1648.s

Harwell, M. C. and Orth, R. J. (2002) Long-Distance Dispersal Potential in a Marine Macrophyte. *Ecology*, 83: 3319-3329.

Hassan, A. N. (2006) *WHO-TDR-SGS*, Final report.

Hazelwood, R.R. (1987). Analyzing the Rape and Profiling the Offender. In: R. R. Hazelwood & A. W. Burgess (eds.), *Practical Aspects of Rape Investigation: A Multidisciplinary Approach.* New York, NY: Elsevier.

Hengeveld, R. (1989) *The dynamics of Biological Invasions.* Oxford University

Press, London.

Henshaw, R. E. and Stephenson, R. O. (1974) Homing in the gray wolf (*Canis lupus*). *Journal of Mammalogy*, 55:234-237.

Hicks, D. W. and Tunnell, J.W.J. (1995) Ecological notes and patterns of dispersal in the recently introduced mussel, *Perna perna* (linne, 1758), in the Gulf of Mexico. *American Malacological Bulletin*, 11:203-206.

Higgins, S. I., & Richardson, D. M. (1996) A review of models of alien plant spread. *Ecological modelling*, 87:249-265.

Higgins, S. I., Nathan, R., & Cain, M. L. (2003). Are long-distance dispersal events in plants usually caused by nonstandard means of dispersal? *Ecology*, 84(8):1945-1956.

Hill, M. *et al*. (2005) Audit of non-native species in England. English Nature Report, (www.english-nature.org.uk).

Holroyd, G.L., Conway, C.J. and Trefry, H.E. (2011) Breeding Dispersal of a Burrowing Owl from Arizona to Saskatchewan. *The Wilson Journal of Ornithology*, 123:378-381.

Huelsenbeck, J.P., Andolfatto, P. (2007) Inference of population structure under a Dirichlet process model. *Genetics*, 175:1787-1802.

Huelsenbeck, J.P., Jain, S., Frost, S. W. D., Kosakovsky Pond, S. L. (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proceedings of the National Academy of Sciences*, 103:6263-6268.

Hulme, P. (2006) Beyond control: wider implications for the management of

biological invasions. *Journal of Applied Ecology*, 43:835–847.

Hulme, P. (2007) Biological invasions in Europe: drivers, pressures, states, impacts and responses. *Issues in Environmental Science and Technology:*56.

Hungerford, K. E. and Wilder, N.G. (1941) Observations on the homing behavior of the gray squirrel (*Sciurus carolinensis*). *Journal of Wildlife Management*, 5:458-460.

J

Jenkins, P. T. (1996) Free trade and exotic species introductions. *Conservation Biology*, 10(1):300-302.

Jones, W.E. and Barb, M.S. (1968) The motile period of swarmers *of Enteromorpha intestinalis* (L.) Link. *British Phycological Bulletin*, 3:525-528.

Jonsen, I. D., Bourchier, R. S. & Roland, J. (2001) The influence of matrix habitat on aphthona flea beetle immigration to leafy spurge patches. *Oecologia*; 127:287-294.

Jonsen, I. D., Myers, R. A., & Flemming, J. M. (2003) Meta-analysis of animal movement using state-space models. *Ecology*; 84:3055-3063.

Jordan J. & Horsburgh, N. (2005) Mapping jihadist terrorism in Spain. *Studies in Conflict and Terrorism*,28:169-191.

K

Kaldy, J. E. and Dunton, K. H. (1999) Ontogenetic photosynthetic changes, dispersal and survival of *Thalassia testudinum* (turtle grass) seedlings in a sub-tropical lagoon. *Journal of Experimental Marine Biology and Ecology*, 240:193-212.

Keith, L. B., and J. D. Waring (1956) Evidence of orientation and homing in snowshoe hares. *Canadian Journal of Zoology*, 34:579-581.

Keller, R. P., & Drake, J. M. (2009) Trait based risk assessment for invasive species. *Bioeconomics of Invasive Species: Integrating Ecology, Economics, Policy and Management*, 44-62.

Keller, R.P., Frang, K. & Lodge, D.M. (2008) Preventing the spread of invasive species: economic benefits of intervention guided by ecological predictions. *Conservation Biology*, 22: 80-88.

Keller, R. P., Lodge, D. M., Lewis, M. A. and Shogren, J. F. (2009*) The Bioeconomics of invasive specie*s. Oxford University Press, Oxford.

Kenchington, E., Duggan, R., and Riddell, T. (1998) Early life history characteristics of the razor clam (*Ensis directus*) and the moonsnails (*Euspira spp.*) with applications to fisheries and aquaculture. Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada.

Keough, M. J. and Chernoff, H. (1987) Dispersal and population variation in the bryozoan Bugula neritina. *Ecology,* 68:199-210.

Kind, S.S. (1987) Navigational ideas and the Yorkshire Ripper investigation. *Journal of Navigation,* 40:385-393.

Klugman, K. P. (1990) Pneumococcal resistance to antibiotics. *Clinical*

microbiology reviews, *3*(2):171-196.

Knowlton, N. and Keller, B.D. (1986) Larvae which fall far short of their potential: highly localized recruitment in an *Alpheid* shrimp with extended larval development. *Bulletin of Marine Science*, 39:213-223.

Kolar, C. S., & Lodge, D. M. (2002) Ecological predictions and risk assessment for alien fishes in North America. *Science*, 298(5596): 1233-1236.

Kot, M., Lewis, M. A., & van den Driessche, P. (1996) Dispersal data and the spread of invading organisms. Ecology, 77(7):2027-2042.

Kucera H. (2005) Hunting insurgents - geographic profiling adds a new weapon. *Geo World,* 30-32.

Kuhn, H. W. and Kuenne, R. E. (1962) An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics. *Journal of Regional Science*, 4:21-33.

L

Laverty, I. and MacLaren, P. (2002) Geographic Profiling: a New Tool for Crime Analysts. *Crime Mapping News*, 4(3):5-8.

LeBeau, J. L. (1992) Four Case Studies Illustrating the Spatial-Temporal Analysis of Serial Rapists. *Police Studies,* 15:124-145.

Lechleitner, R. R. (1958) Movements, density, and mortality in a black-tailed jackrabbit population. *Journal of Wildlife Management*, 22:371-384.

Le Comber, S. C. & Stevenson, M. D. (2012) From Jack the Ripper to epidemiology and ecology. *Trends In Ecology and Evolution*, 27, 307.

Le Comber SC, Nicholls B, Rossmo DK, Racey PA.(2006) Geographic profiling and animal foraging. *Journal of Theoretical Biology* 240:233-240.

Le Comber SC, Rossmo DK, Hassan AN, Fuller DO, Beier JC.(2011) Geographic profiling as a novel spatial tool for targeting infectious disease control. *International Journal of Health Geographics* 10:35.

Lendrum, P. E., Elbroch, L. M., Quigley, H., Thompson, D. J., Jimenez, M., & Craighead, D. (2014) Home rage characteristics of a subordinate predator: selection for refugia or hunt opportunity? *Journal of Zoology* 294:58-66.

Leung, B., Lodge, D.M., Finnoff, D., Shogren, J.F., Lewis, M.A. & Lamberti, G. (2002) An ounce of prevention or a pound of cure: bioeconomic risk analysis of invasive species. *Proceedings of the Royal Society of London Series B: Biological Sciences* 269:2407–2413.

Leung, B., Drake, J. and Lodge, D. (2004) Predicting invasions: propagule pressure and the gravity of Allee effects. *Ecology* 85:1651-1660.

Levine, N. (2009) *CrimeStat: A spatial statistics program for the analysis of crime incident locations*. Ned Levine & Associates, Annandale, VA and the National Institute of Justice, Washington, DC. (http://www.icpsr.umich.edu/crimestat).

Lewis, M. A., & Kareiva, P. (1993) Allee dynamics and the spread of invading organisms. *Theoretical Population Biology* 43:141-158.

Link, W. A., Eaton, M. J. (2012) On thinning chains in MCMC. *Methods in Ecology*

and Evolution, 3: 112-115.

Linkhart, B.D. and Reynolds, R.T. (2007) Return rate, fidelity, and dispersal in a breeding population of flammulated owls *(Otus flammeolus). The Auk*, 124(1):264-275.

Lockwood, J. L., Cassey, P., & Blackburn, T. (2005) The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution*, 20(5):223-228.

Lodge D. M. (1993) Biological invasions: lessons for ecology. *Trends in Ecology & Evolution*, 8:133–137.

Loecher, M. (2012) *RgoogleMaps: Overlays on Google map tiles in R*. R package version 1.2.0.2. Berlin School of Economics and Law. <http://CRAN.Rproject.org/package=RgoogleMaps>

Logan, K. A., and Sweanor, L.L. (2000) *Ecology and management of large mammals in North America. Puma.* pp. 347-377 in S. Demarais and P. Krausman, (ed). Prentice-Hall, Englewood.

M

Mack, A. L. (1995) Distance and non-randomness of seed dispersal by the dwarf cassowary Casuarius bennetti. *Ecography*, 18(3): 286-295.

Magdziarz, M., & Teuerle, M. (2015) Asymptotic properties and numerical simulation of multidimensional Levy walks. *Communications in Nonlinear Science and Numerical Simulation*, 20:489-505.

Marliave, J. B. (1986) Lack of planktonic dispersal of rocky intertidal fish larvae. *Transactions of the American Fisheries Society*, 115:149-154.

Martin RA, Rossmo DK, Hammerschlag N. (2009) Hunting patterns and geographic profiling of white shark predation. *Journal of Zoology*, 279:111-118.

MacIsaac, H. J., Borbely, J. V., Muirhead, J. R., & Graniero, P. A. (2004) Backcasting and forecasting biological invasions of inland lakes. *Ecological Applications*, 14(3):773-783.

May, R. M. (1981) Theoretical ecology. Principles and applications. *Theoretical ecology. Principles and applications* (second edition).

McDermott, J. J. (1998) The western Pacific brachyuran (*Hemigrapsus sanguineus: Grapsidae*), in its new habitat along the Atlantic coast of the United States: geographic distribution and ecology. *International Council for Exploration of the Sea Journal of Marine Science*, 55:289-298.

McIntosh, R.P. (1985) *The background of ecology*. Cambridge University Press, Cambridge, UK.

Meinesz, A., de Vaugelas, J., Hesse, B. and Mari, X. (1993) Spread of the introduced tropical alga *Caulerpa taxifolia* in northern Mediterranean waters. *Journal of Applied Phycology,* 5:141-147.

Miller, T. D. (1996) First record of the green crab*, Carcinus maenas*, in Humboldt Bay, California. *California Fish Game Bulletin*, 82:93-96.

Miller, S. D. and Ballard, W.B. (1982) Homing of transplanted Alaskan brown bears. *Journal of Wildlife Management*, 46:869-876.

Moilanen, A., and Hanski, I. (1998) Metapopulation dynamics: effects of habitat quality and landscape structure. *Ecology* 69:143-153.

Mooney H.A., Drake J.A., eds. (1986) *Ecology of Biological Invasions of North America and Hawaii*. New York: Springer-Verlag.

Mooney, H. A., & Hobbs, R. J. (2000) Global change and invasive species: where do we go from here. *Invasive Species in a Changing World*. Washington, DC: Island, 425-434.

Moorcroft, P. R., Lewis, M. A., & Crabtree, R. L. (1999) Home range analysis using a mechanistic home range model. *Ecology*; 80:1656-1665.

Morton, M L. (1997) Natal and breeding dispersal in the mountain white crowned sparrow *Zonotrichia leucophrys oriantha*. *Ardea*, 85:145-154.

Murie, 0. J. & Murie, A. (1931) Travels of *Peromyscus*. *Journal of Mammalogy*, 12: 200-209.

N

Narbona,, E., Arista, M. and Ortiz, P.L. (2005) Explosive seed dispersal in two perennial Mediterranean Euphorbia Species (*Euphorbiaceae*). *American Journal of Botany,* 92:510-517.

Nathan, R., Muller-Landau, H.C. (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends In Ecology & Evolution*, 15:278-285.

Nathan, R. (2001) The challenges of studying dispersal. *Trends in Ecology and Evolution*,16:481-483.

Neal, R.M. (2000) Markov chain sampling methods for Dirichlet Process mixture models. *Journal of computational and graphical statistics*, 9:249-265.

Nekola, J. C. and White P.S. (1999) Special Paper: The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography*, 26(4):867-878.

Newton, Jr., M.B. & D.C. Newton (1985) *Geoforensic identification of localized serial crime: unsolved female homicides*, Fort Worth, Texas, 1983-85. Paper presented at the meeting of the Southwest Division, Association of American Geographers.

Nicholson, K. L., Ballard, W. B., McGee, B. K., & Whitlaw, H. A. (2007) Dispersal and extraterritorial movements of swift foxes (*Vulpes velox*) in northwestern Texas. *Western North American Naturalist*, 67(1):102-108.

Nikaido, H. (1998) Antibiotic resistance caused by gram-negative multidrug efflux pumps. *Clinical Infectious Diseases*, 27:32-41.

Nocedal, J. and Wright, S. J. (1999) *Numerical Optimization*. Springer.

Noel, P. Y., Tardy, E. and D'Udekem D'Acoz, C. (1997) Will the crab *Hemigrapsus penicillatus* invade the coasts of Europe? Les Comptes Rendus de l'Academie des Sciences (Paris) 320:741-745.

O

O'Leary, M. (2009) The mathematics of geographic profiling. *Journal of Investigative Psychology and Offender Profiling*, 6:253-265.

O'Leary, M. (2010a) Multimodel Inference and Geographic Profiling. *Crime Mapping*, 2(1).

O'Leary, M. (2010b) Implementing a Bayesian approach to criminal geographic profiling. *First International Conference on Computing for Geospatial Research and Application*, June 21-23 Washington, D.C.

O'Leary, M. (2012) *New Mathematical Approach to Geographic Profiling*, National Insitute of Justice, Washington, D.C.

Olson, R. R. (1983) Ascidian-Prochloron symbiosis: the role of larval photoadaptations in midday larval release and settlement. *Biological Bulletin,* 165:221-240.

Olson, R. R. and McPherson, R. (1987) Potential vs. realized larval dispersal: fish predation on larvae of the ascidian *Lissoclinum patella* (Gottschaldt). *Journal of Experimental Marine Biology and Ecology*, 110:245-256.

Ordinance Survey (2010) *A guide to coordinate systems in Great Britain*. Ordinance Survey, London viewed 01 July 2013, <http://www.ordnancesurvey.co.uk/oswebsite/docs/support/guide-coordinate-systems-great-britain.pdf>

Orr, R. (2003) Generic nonindigenous aquatic organisms risk review process. pp 415-438 in G.M. Ruiz and T.Carlton, (ed). *Invasive species, vectors and management strategies*. Island Press, Washington.

Osorio-Beristain, M. and Drummond, H. (1993) Natal dispersal and deferred breeding in the blue-footed booby. *The Auk,* 110(2):234-239.

Ostfeld, R. S. and Manson R.H. (1996) Long-distance homing in meadow voles, *Microtus pennsylvanicus*. *Journal of Mammalogy*, 77:870-873.

Othmer, H. G., Dunbar, S. R., & Alt, W. (1988) Models of dispersal in biological systems. *Journal of mathematical biology*, 26(3):263-298.

P

Paradis, E., Baillie, S. R., Sutherland, W. J., & Gregory, R. D. (1998) Patterns of natal and breeding dispersal in birds. *Journal of Animal Ecology*, 67(4):518-536.

Paterson, DL, 2006. Resistance in Gram-Negative Bacteria: Enterobacteriaceae. *The American Journal of Medicine*, 119 (6):S20-S28

Papini, A., Mosti, S., & Santosuosso, U. (2012). Tracking the origin of the invading *Caulerpa (Caulerpales, Chlorophyta*) with Geographic Profiling, a criminological technique for a killer alga. *Biological Invasions*, 1-9.

Payne, N. F (1975) Unusual movements of Newfoundland black bears. *Journal of Wildlife Management*, 39:812-813.

Pebesma, E.J., & R.S. Bivand, (2005). Classes and methods for spatial data in R. *R. News*, <http://cran.r-project.org/doc/Rnews/>

Peters, R.H. (1991) *A critique for ecology*. Cambridge University Press, Cambridge, UK.

Peterson A. T. (2003) Predicting the geography of species invasions via ecological niche modelling. *The quarterly review of biology,* 78(4):419-433.

Phillips, R. L. and Mech, L.D. (1970) Homing behavior of a red fox. *Journal of Mammalogy*, 51:621.

Pielowski, Z. (1972) Home range and degree of residence of the European hare. *Acta Theriologica*, 9:93-103.

Pimentel, D., S. McNair, J. Janecka, J. Wightman, C. Simmonds, C. O'Connell, E. Wong, L. Russel, J. Zern, T. Aquino and T. Tsomondo (2001) Economic and environmental threats of alien plant, animal, and microbe invasions. *Agriculture, Ecosystems & Environment.* 84(1):1-20.

Powers, K.S. and Aviles, L. (2003) Natal Dispersal Patterns of a Subsocial Spider *Anelosimus cf. jucundus (Theridiidae*). *Ethology*, 109:725-737.

Prince, J. D., Sellers, T.L., Ford, W.B. and Talbot, S.R. (1987) Experimental evidence for limited dispersal of haliotid larvae (genus *Haliotis*; Mollusca: Gastropoda). *Journal Experimental Marine Biology and Ecology*, 106:243-263.

Puth, L.M. & Post, D.M. (2005) Studying invasion: have we missed the boat? *Ecology Letters*, 8:715-721.

Pysek, P. (1995) On the terminology used in plant invasion studies. *Plant invasions: general aspects and special problems* . P. Pysek, K. Prach, M. Rejm·nek and M. Wade (eds), pp. 71-81. SPB. Academic Publishing, Amsterdam, NL.

Pyšek, P., Jarošík, V., Hulme, P. E., Pergl, J., Hejda, M., Schaffner, U., & Vilà, M. (2012) A global assessment of invasive plant impacts on resident species,

communities and ecosystems: the interaction of impact measures, invading species' traits and environment. *Global Change Biology*, 18(5):1725-1737.

R

R Development Core Team (2012) R: A language and environment for statistical computing. *R Foundation for Statistical Computing* , <www. R-project.org>.

Raine, N. E., Rossmo, D. K., Le Comber, S. C. (2009) Geographic profiling applied to testing models of bumble-bee foraging. *Journal of the Royal Society Interface*, 6:307–319.

Randall, J. E. (1987) Introductions of marine fishes to the Hawaiian Islands. *Bulletin of Marine Science*, 41:490-502.

Randall, J.E., Earle, J.L., Pyle, R.L., Parrish, J.D. and Hayes, T. (1993) Annotated checklist of the fishes of Midway Atoll, Northwestern Hawaiian Islands. *Pacific Science,* 47: 356-400.

Reed, D. C., Laur, D.R. and Ebeling, A.W. (1988) Variation in algal dispersal and recruitment: the importance of episodic events. *Ecological Monographs*, 58: 321-335.

Rejmnek, M., Richardson, D.M., Barbour, M.G., Crawley, M.J.,Hrusa, G.F., Moyle, P.B., Randall, J.M., Simberloff, D. & Williamson, M. (2002) Biological invasions: politics and the discontinuity of ecological terminology. *ESA Bulletin*, 83:131-133.

Reynolds, A. M. (2014) Detecting Lévy walks without turn designation. *Behavioral*

Ecology and Sociobiology, 68:1893-1899

Ricciardi, A. (2007) Are modern biological invasions an unprecedented form of global change? *Conservation Biology*, 21:329–336.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832-837.

Rossmo, D. K. (1993) A methodological model. *American Journal of Criminological Justice*, 172: 1–21.

Rossmo, D.K . (1995) *Geographic profiling: Target patterns of serial murderers*, Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC, Canada.

Rossmo, D.K., (2000) *Geographic profiling.* CRC Press, Boca Raton, Florida, USA.

Rossmo, D. K. (2012) Recent developments in geographic profiling. *Policing,* 6:144-150.

Rossmo, D. K. and Harries, K. D. (2011) The geospatial structure of terrorist cells. *Justice Quarterly*, 28:221–248.

Rossmo, D. K. and Velarde, L. (2008) Geographic profiling analysis: principles, methods, and applications. In Chainey, S. and Tompson, L (eds), *Crime Mapping Case Studies: Practice and Research*. Chichester: John Wiley & Sons, pp. 35–43.

Rossmo, D. K., Davies, A. and Patrick, M. (2004) *Exploring the geo-demographic and distance relationships between stranger rapists and their offences* (Special Interest Series: Paper 16). London: Research, Development and Statistics Directorate, Home Office.

Roy, K. D. (1991) Cited in R. A. Powell (1993) *The fisher: life history, ecology and*

behaviour. Second edition. University of Minnesota Press, Minneapolis, Minnesota, USA.

Ruth, T. K., Logan, K. A., Sweanor, L. L., Hornocker, M. G. and Temple, L. J. (1998) Evaluating cougar translocation in New Mexico. *Journal of Wildlife Management,* 62:1264-1275.

Rutherglen, R. A. and Herbison, B. (1977) Movements of nuisance black bears (*Ursus americanus*) in southeastern British Columbia. *Canadian Field-Naturalist*, 91:419-422.

S

Saitoh, T. (1995) Sexual Differences in Natal Dispersal and Philopathy of the Grey-Sided Vole. *Researches on Population Ecology*, 37(1):49-57.

Sakai, A K., Allendorf, F W., Holt, J. S., Lodge, D M., Molofsky, K. A. (2001) The Population Biology of Invasive Species. *Annual Review of Ecology and Systematics* 32:305-332.

Sammarco, P. W. and Andrews, J.C. (1989) The helix experiment: differential localized dispersal and recruitment patterns in Great Barrier Reef corals. *Limnology and Oceanography* 34:896-912.

Sapp, A.D., Huff, T.G., Gary, G.P., Icove, D.J., Horbert, P. (1994) *A report of essential findings from a study of serial arsonists*. National Center for the Analysis of Violent Crime: Quantico, VA

Saville, N. M., Dramstad, W. E., Fry, G. L., & Corbet, S. A. (1997) Bumblebee

movement in a fragmented agricultural landscape. *Agriculture, ecosystems & environment* 61(2):145-154.

Scheltema, R. S. (1971) Larval dispersal as a means of genetic exchange between geographically separated populations of shoalwater benthic marine gastropods. *Biological Bulletin* 140:284-322.

Schito, G. C., Debbia, E. A., & Marchese, A. (2000) The evolving threat of antibiotic resistance in Europe: new data from the Alexander Project. *Journal of Antimicrobial Chemotherapy* 46(3):3-9.

Seidel, D. R. (1961) Homing in the eastern chipmunk. *Journal of Mammalogy* 42:256-257.

Serrano, D., Tella, J. L., Donzar, J. A., Pomarol, M. (2003) Social and individual features affecting natal dispersal in the colonial lesser kesterel (2003) *Ecology* 84(11):3044-3054.

Singh, H.P., Batish, D.R., Kohil, R.K. (2001) Allelopathy in agroecostystems. *Journal of crop production* 4(2):1-41.

Slough, B. G. (1989) Movements and habitat use by transplanted marten in the Yukon Territory. *Journal of Wildlife Management* 53:991-997.

Smith, A. J. (1975) Invasion and ecesis of bird-disseminated woody plants in a temperate forest sere. *Ecology* 19-34.

Smith, M. A., & Green, D. M. (2005) Dispersal and the metapopulation paradigm in amphibian ecology and conservation: are all amphibian populations ,metapoplations? *Ecography* 28:110-128.

Snow J. & Frost W. (1936) *Snow on cholera*: Being a reprint of two papers. Commonwealth.

Sovada, M.A, Slivinski, C.C., Woodward, R.O. and Phillips, M.L. (2003) Home range, habitat use, litter size, and pup dispersal of swift foxes in two distinct landscapes of western Kansas. pp 149–160 in M.A Sovada and L. Carbyn, (eds), *The swift foxes: ecology and conservation of swift foxes in a changing world*. Canadian Plains Research Center, Regina, Saskatchewan, Canada.

Starfinger, U. (1998) On success in plant invasions. *Starfinger, U.; Edwards, K.; Kowarik* (eds.) I:33-42.

Stevenson, P. R. (2000) Seed dispersal by woolly monkeys (Lagothrix lagothricha) at Tinigua National Park, Colombia: Dispersal distance, germination rates, and dispersal quantity. *American Journal of Primatology* 50(4):275-289.

Stoner, D. S. (1990) Recruitment of a tropical colonial ascidian: relative importance of re-settlement vs. post-settlement processes. *Ecology* 71:1682-1690.

Stoner, D. S. (1992) Vertical distribution of colonial ascidian on a coral reef: the roles of larval dispersal and life history variation. *American Naturalist* 139:802-824

Stevenson, M. D., Rossmo, D. K., Knell, R. J., Le Comber, S. C. (2012) Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography* 10:704-715.

Strayer, D.L., Evinver, V.T., Jeschke, J.M. & Pace, M.L. (2006) Understanding the long-term effects of species invasions. *Trends in Ecology & Evolution* 21:645-651.

Sweanor, L.L., Logan, K.A. and Hornocker, M.G. (2000) Cougar Dispersal Patterns,

Metapopulation Dynamics, and Conservation. *Conservation Biology* 14:798-808.

T

Tabatabai, F. R., & Kennedy, M. L. (1989) Movements of relocated raccoons (*Procyon lotor*) in western Tennessee. *Journal of the Tennessee Academy of Science* 64:221-224.

Thorson, G. (1946) Reproduction and larval development of Danish marine bottom invertebrates, with special reference to the planktonic larvae in the Sound (Oresund). *Meddelelser fra Danmarks fiskeri- og havundersøgelser. Ser Plankton* 4:1-523.

Thuiller, W., Richardson, D.M., Pysek, P., Midgley, G.F., Hughes, G.O. & Rouget, M. (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology* 11:2234–2250.

Townes, H. (1962) Design for a Malaise trap. *Proceeding of the Entomological Society Washington* 64(4):253-262.

Trakhtenbrot, A., Nathan, R., Perry, G., & Richardson, D. M. (2005) The importance of long distance dispersal in biodiversity conservation. *Diversity and Distributions* 11(2):173-181.

Truve, J. and Lemel, J. (2003) Timing and distance of natal dispersal for wild boar *Sus scrofa* in Sweden. *Wildlife Biology* 9 (1):51-57.

V

Vandermeulen, H. and DeWreede, R.E. (1986) The phenology, mortality, dispersal and conopy species interaction of *Colpomenia peregrina* (Sauv.) Hamel in British Columbia. *Journal Experimental Marine Biology and Ecology* 99:31-47.

Van Houtan, K.S., Pimm, S. L., Halley, J. M., Bierregaard, R. O. Jr., Lovejoy, T. E. (2007) Dispersal of Amazonian birds in continuous and fragmented forest. *Ecology Letters* 10: 219-229.

Vaughan, T. A. (1963) Movements made by two species of pocket gophers. *American Midland Naturalist* 69:367-371.

Venable, D. L. and Lawlor, L. (1980) Delayed germination and dispersal in desert annuals: escape in space and time. *Oecologia* 46:272-282.

Venable, D.L., Flores-Martinez, A., Muller-Landau, H.C., Barron-Gafford, G. and Becerra, J.X. (2008) Seed Dispersal of Desert Annuals. *Ecology* 89:2218-2230.

Verma, A., & Lodha, S. K. (2002) A topological representation of the criminal event. *Western Criminology Review* 3(2):1-30.

Vermeij, G. J. (1978) *Biogeography and adaptation*. Harvard University Press, Cambridge, Massachusetts, USA.

Vitousek, P. M., D'Antonio, C. M., Loope, L. L., & Westbrooks, R. (1996) Biological invasions as global environmental change. *American Scientist* 84(5):468-478.

Vitousek, P. M., Mooney, H. A., Lubchenco, J. and. Melillo, J. M (1997) Human Domination of Earth's Ecosystems. *Science* 277(5325):494.

W

Walker, K., & Lynch, M. (2007) Contributions of Anopheles larval control to malaria suppression in tropical Africa: review of achievements and potential. *Medical and veterinary entomology* 21(1):2-21.

Wallace, A. R. (1880) *Island Life*, MacMillan, London.

Ward, R. G. and Brookfield, M. (1992) The dispersal of the coconut: did it float or was it carried to Panama. *Journal of Biogeography* 19:467-480.

Warren, P. K. and Baines D. (2002) Dispersal, survival and causes of mortality in black grouse *Tetrao tetrix* in northern England. *WildlifeBiology* 8:91-97.

Weins, J. D., Reynolds, R. T. and Noon, B. R. (2006) Juvenile movement and natal dispersal of northern goshwaks in Arizona. *The Condor* 108:253-269.

Weiszfeld, E. (1936) Sur le point pour lequel la somme des distances den points donnes est minimum. *Tohoku Mathematical Journal* 43:355-386.

Wheelwright, N. T., Mauck, R.A. (1998) Philopatry, natal dispersal and inbreeding avoidance in an island population of savannah sparrows. *Ecology* 79(3):755-767.

Wiens, J. J., & Graham, C. H. (2005) Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual review of ecology, evolution, and systematics* 519-539.

Wilcove, D. S Rothstein, D., Dubow, J., Phillips A and Losos E (1998) Quantifying threats to imperiled species in the United States. *BioScience* 48(8):607-615.

Wilkinson, J. W. (2011) Alpine Newt, *Mesotriton alpestris*.[Online].Avaliable at <https://secure.fera.defra.gov.uk/nonnativespecies/factsheet/factsheet.cfm?speciesId =2215 > Accessed: 14.2.2013.

Williamson, M. & Fitter, A. (1996) The varying success of invaders. *Ecology* 77:1661–1666.

Wilson, E. 0. (1992) *The Diversity of Life*. New York: Norton and Company.

Winkler, D. W., Wrege, P. W., Alllen, P. E., Kast, T. L., Senesac, P., Wasson, M. F., Sullivan, P. J. (2005) The natal dispersal of tree swallows in a continuous mainland environment. *Journal of Animal Ecology* 74:1080–1090.

Worcester, S. E. (1994) Adult rafting versus larval swimming: dispersal and recruitment of a botryllid ascidian on eelgrass. *Marine Biology* 121:309-318.

Worton, B. J. (1989) Kernel methods for estimating the utilization distribution in home range studies. *Ecology* 70:164:168

Y

Yardena, S. I., Boaz, F., Ruth, O. W., Yoav, G., Aliza, N., David, S., & Michael, G. (2002) Reappraisal of community-acquired bacteremia: a proposal of a new classification for the spectrum of acquisition of bacteremia. *Clinical infectious diseases 34*(11):1431-1439.

Yohannes M, Haile M, Ghebreyesus T, Witten K, Getachew A, Byass P, Lindsay, S. (2005) Can source reduction of mosquito larval habitat reduce malaria transmission in Tigray, Ethiopia? Tropical *Medicine and International Health* 10:1274–1285.

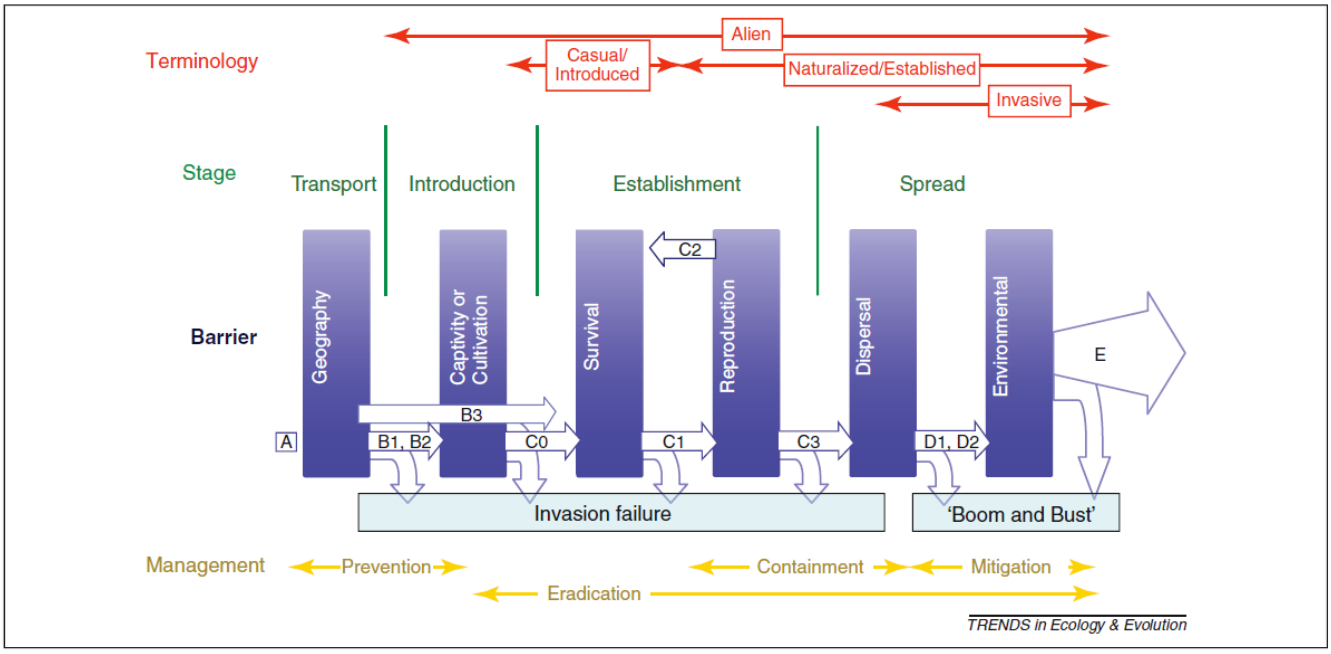
Z

Zechman, F. W., & Mathieson, A. C. (1985). The distribution of seaweed propagules in estuarine, coastal and offshore waters of New Hampshire, USA. *Botanica Marina* 28(7):283-294.

Zimmermann, F., Breitenmoser Würsten, C., & Breitenmoser, U. (2005) Natal dispersal of the Eurasian lynx in Switzerland. *Journal of the Zoological Society of London* 267:381-395.

Zipf, G. (1950) *The principle of least effort* (Addison Wesley, Reading, MA).

Appendix A: A unified invasion framework



The unfied invasion framework taken from Blackburn *et al.* (2011), in which the invasion is shown as a series of stages; within each stage there are barriers that must be overcome in order to pass to the next. For a full description of this framework see Blackburn *et al.* (2011)

Appendix B: R code for the original DPM model

```
####    GP_DirichletProcess_RgoogleMaps.R


#       Description:

#       Similar to GP_DirichletProcess2.R, except uses input data rather than simulated data.

#       Produces Googlemaps output


#       Author: Robert Verity, Mark Stevenson

#       Date: 22/08/2012


# DECLARE FUNCTIONS AND INSTALL NECCESSARY PACKAGES #


install.packages("RgoogleMaps")

install.packages("rgdal")

install.packages("ggplot2")


library(RgoogleMaps)

library(rgdal)

library(ggplot2)

library(grDevices)


#### Get map based on data OR specified zoom window

GetMapClever <-
function(datax=NULL,datay=NULL,WindowLong=NULL,WindowLat=NULL,gridsize=640,maptype="roadmap",destfile=destfile) {

        # Set zoom by data
```

```
	if (length(WindowLong)==0) {

		zoom =
max(MaxZoom(c(min(datay),max(datay)),c(min(datax),max(datax)),size=c(gridsize,
gridsize)))

		center = c((min(datay)+max(datay))/2,(min(datax)+max(datax))/2)

		MyMap =
GetMap(center=center,size=c(gridsize,gridsize),zoom=zoom,maptype=maptype,dest
file=destfile)

		}

	# Set zoom by window

	if (length(WindowLong)>0) {

		zoom =
max(MaxZoom(WindowLat,WindowLong,size=c(gridsize,gridsize)))

		center = c(sum(WindowLat)/2,sum(WindowLong)/2)

		MyMap =
GetMap(center=center,size=c(gridsize,gridsize),zoom=zoom,maptype=maptype,dest
file=destfile)

		}

	return(MyMap)

	}


#### Plot map based on saved map dimensions OR specified zoom window

PlotMap <- function(MyMap,WindowLong=NULL,WindowLat=NULL) {

	# Set zoom by data

	if (length(WindowLong)==0) {

		xwindow=c(MyMap[5]$BBOX$ll[2],MyMap[5]$BBOX$ur[2])

		ywindow=c(MyMap[5]$BBOX$ll[1],MyMap[5]$BBOX$ur[1])


	plot(1,type="n",xlim=xwindow,ylim=ywindow,xaxs="i",yaxs="i",xlab="Lon
gitude",ylab="Latitude")
```

```
        rasterImage(MyMap$myTile,xwindow[1],ywindow[1],xwindow[2],ywindow
[2])
                }
        # Set zoom by window
        if (length(WindowLong)>0) {

        plot(1,type="n",xlim=WindowLong,ylim=WindowLat,xaxs="i",yaxs="i",xla
b="Longitude",ylab="Latitude")

        rasterImage(MyMap$myTile,MyMap[5]$BBOX$ll[2],MyMap[5]$BBOX$ll[
1],MyMap[5]$BBOX$ur[2],MyMap[5]$BBOX$ur[1])
                }
        }

#### Convert matrix to translucent raster image
CreateRaster <- function(matrix,levels,transp) {
        tempmat = matrix(1,nrow(matrix),ncol(matrix))
        for (i in 1:(length(levels))) {
                tempmat = tempmat + ((matrix/max(matrix))>levels[i])
                }
        colvec <<- c("transparent",heat.colors(length(levels)-1))
        transp.bit = round(transp*255)
        transp.string = as.hexmode(transp.bit)
        if (transp.bit<16) {transp.string=paste("0",transp.string,sep="")}
        for (i in 2:length(colvec)) {
                colvec[i] = paste(substr(colvec[i],1,7),transp.string,sep="")
                }
        outmat = matrix(colvec[tempmat],nrow=nrow(matrix))
```

```
	return(outmat)
	}

#### Display contours that work for any zoom level
Contours <- function(xvec,yvec,matrix,levels) {
	flipmat = t(matrix[nrow(matrix):1,])
	conts = contourLines(xvec,yvec,flipmat,levels=levels)
	for (i in 1:length(conts)) {
		lines(conts[[i]]$x,conts[[i]]$y,col="dark grey")
		}
	}

#### Plot default Google map with surface and points
DefaultMap <-
function(MyMap,xvec=NULL,yvec=NULL,Surface=NULL,levels=NULL,transp=NULL,data=NULL,msources=NULL) {
	PlotMap(MyMap)
	if (length(Surface)>0 & length(levels)>0 & length(transp)>0 & length(xvec)>0 & length(yvec)>0) {

	rasterImage(CreateRaster(Surface,levels,transp),MyMap$BBOX$ll[2],MyMap$BBOX$ll[1],MyMap$BBOX$ur[2],MyMap$BBOX$ur[1])
		Contours(xvec,yvec,Surface,levels)
		}
	if (length(data)>0) {
		points(data[,1],data[,2],pch=20,cex=0.8)
		}
	if (length(msources)>0) {
		points(msources[,1],msources[,2],pch=15,col=4)
```

```
        }
    }

#### As above with zoom capability

ZoomMap <-
function(MyMap,xvec=NULL,yvec=NULL,Surface=NULL,levels=NULL,transp=NULL,data=NULL,msources=NULL,maptype=MapType) {

    DefaultMap(MyMap,xvec=xvec,yvec=yvec,Surface=Surface,levels=levels,transp=0.4,data=data,msources=msources)

    ChooseWindow = locator(2)

    MyMap2 =
GetMapClever(WindowLong=sort(ChooseWindow$x),WindowLat=sort(ChooseWindow$y),destfile=paste(Location,"test.png",sep=""),maptype=maptype)

    PlotMap(MyMap2,WindowLong=sort(ChooseWindow$x),WindowLat=sort(ChooseWindow$y))

    if (length(Surface)>0 & length(levels)>0 & length(transp)>0 &
length(xvec)>0 & length(yvec)>0) {


    rasterImage(CreateRaster(Surface,levels,transp),MyMap$BBOX$ll[2],MyMap$BBOX$ll[1],MyMap$BBOX$ur[2],MyMap$BBOX$ur[1])

        Contours(xvec,yvec,Surface,levels)

        }

    if (length(data)>0) {

        points(data[,1],data[,2],pch=20,cex=0.8)

        }

    if (length(msources)>0) {

        points(msources[,1],msources[,2],pch=15,col=4)

        }

    }
```

```
#### Compute pairwise distances between data
Pairwise <- function(data) {
        xmat1 = outer(rep(1,n),data[,1])
        xmat2 = outer(data[,1],rep(1,n))
        xdist = abs(xmat1-xmat2)
        ymat1 = outer(rep(1,n),data[,2])
        ymat2 = outer(data[,2],rep(1,n))
        ydist = abs(ymat1-ymat2)
        zdist = sqrt(xdist^2+ydist^2)
        output = zdist[col(xmat1)>row(xmat1)]
        return(output)
        }

#### Import data from .txt or .csv adaptively
ImportData <- function(header=F) {
        filepath = tryCatch(file.choose(), error = function(e) NULL)
        if (length(filepath)==0) {
                cat("Import cancelled by user\n")
                output = NULL
        } else {
                extension = tail(unlist(strsplit(filepath,"[.]")),1)
                if (extension=="txt") {output =
as.matrix(read.table(filepath,header=header))}
                if (extension=="csv") {output =
as.matrix(read.csv(filepath,header=header))}
                return(output)
        }
}
```

```
# INPUT STARTING PARAMETERS ----------------------------------------------


#### Setup the colourspace


#### Import data

data = ImportData(header=F)

datax = data[,1]

datay = data[,2]

n = length(datax)


MCMCcols <-
colorRampPalette(c('red','green','orange','blue','yellow','gray','black','brown','aquamar
ine3','cyan','darkmagenta','darkviolet','green4'))

MCMCcols2 = sample(MCMCcols(n))


#### Import sources (optional)

input <- "NA"

while(!isTRUE(input=="Y") && !isTRUE(input=="N")) {

        cat("Import source data? Enter Y=Yes or N=No\n")

        input <- scan("",what="character",n=1,quiet=T)

        if (input=="Y") {

                msources = ImportData(header=F)

                if (length(msources)>0) cat("Source data imported\n")

        } else if (input=="N") {

                cat("Not importing source data\n")

                msources = NULL

        } else {cat("Incorrect input: ")}
```

```
}


#msources=read.table(file.choose())
#othersources=read.table(file.choose())


#### Histogram pairwise distances between data
par(mfrow=c(1,1))
pairwise = Pairwise(data)
hist(pairwise,breaks=2*n,col=8)
abline(v=0.175,col=2)


#### Visualisation parameters


nring=20                    #Number of levels for contour plots
transp=0.4                  #Transparncy of profile overlay
gridsize =      640         #Number of cells in map grid (same in both dimensions) up to 640 max
gridsize2 = 300             #Model resolution
MapType= "roadmap"          #Map type can be any one of several types "roadmap","mobile","satellite", "terrain","hybrid" etc
#Location= "C:\\Documents and Settings\\Mark Stevenson\\My Documents\\Dropbox\\Work Shared\\Bob\\"
#Location= "~/Desktop/Dropbox/Work Shared/Bob/"
Location= "~/Desktop/GP output/"


#### Input model and MCMC parameters
```

```
sigma = 0.175                    #standard deviation of kill profile (Normal
distribution)

tau = "DEFAULT"                          #standard deviation of prior on source
location (Normal distribution). Set as "DEFAULT" for default


minburnin = 100                  #run burnin for at least this long

maxburnin = 250                  #run burnin for at most this long

chains = 5                       #number of chains to run simultaneously

miniterations = 100              #take samples for at least this many iterations

maxiterations = 200              #take samples for at most this many iterations

maxSE = 0.01                     #stop taking samples when this standard error is
reached (and miniterations exceeded)


# PLOT PRIOR ON MAP ------------------------------------------------


#### Download map and extract some useful measures


MyMap =
GetMapClever(datax=datax,datay=datay,destfile=paste(Location,"RawMap.png",sep
=""),maptype=MapType)


#MyMap =
GetMapClever(WindowLong=c(113.8,114.6),WindowLat=c(22.2,23),destfile=paste(
Location,"RawMap.png",sep=""),maptype=MapType)


xmin = MyMap[5]$BBOX$ll[2]

xmax = MyMap[5]$BBOX$ur[2]

ymin = MyMap[5]$BBOX$ll[1]

ymax = MyMap[5]$BBOX$ur[1]
```

```
#### Create prior


priorx = (xmin+xmax)/2

priory = (ymin+ymax)/2


if (tau=="DEFAULT") {
        xdiff = max(datax)-min(datax)
        ydiff = max(datay)-min(datay)
        tau = max(c(xdiff,ydiff))
}


xvec = seq(xmin,xmax,length.out=gridsize2)

yvec = seq(ymin,ymax,length.out=gridsize2)

xmat = outer(rep(1,gridsize2),xvec)

ymat = outer(yvec[gridsize2:1],rep(1,gridsize2))

priormat = dnorm(xmat,mean=priorx,sd=tau)*dnorm(ymat,mean=priory,sd=tau)


#### Plot Google map and overlay prior

levels=seq(0,1,length.out=nring+1)

PlotMap(MyMap)

priorraster = CreateRaster(priormat,levels,transp)

rasterImage(priorraster,xmin,ymin,xmax,ymax)

par(new=T);
contour(xvec,yvec,t(priormat/max(priormat)),xaxs="i",yaxs="i",levels=levels,axes=
F,drawlabels=F,col="dark grey")

points(datax,datay,pch=20,cex=0.8,col="red")
```

```
# INTEGRATION ------------------------------------------------

#### Integrate over hyper-prior on alpha (some fancy integration tricks to make this possible). Output in log space, where the ith element of the vector integrated_prob contains the logged integral of (x^i)*gamma(x)/gamma(n+x) over the hyperprior 1/(1+x)^2

integrated_prob = rep(0,n)

for (i in 1:n) {

	temp = rep(0,1001)

	for (j in 2:1001) {

		integrand = function(x) {

			exp((j-501)*log(10) + i*log(x*n) + lgamma(x*n)-lgamma(n+x*n) -2*log(1+x*n))

			}

		temp[j] = integrate(integrand,lower=0,upper=Inf)$value*n

		if (temp[j]<0) temp[j]=0

		temp[j] = log(temp[j]) - (j-501)*log(10)

		if (temp[j]!=-Inf & abs(temp[j]-temp[j-1])<0.0001) {

			integrated_prob[i] = temp[j]

			break()

			}

		}

	}


#### START MCMC BURNIN LOOP  ---------------------------------------------


#### Initialise MCMC objects

group = t(mapply(sample,size=rep(n,chains),MoreArgs=list(x=n,rep=T)))   #initial group assignment is random. Every row represents a chain, every column represents
```

```
a data point

frequencies = matrix(0,nrow=chains,ncol=n)#initialise frequency matrix

sumfreqs_x = matrix(0,nrow=chains,ncol=n)#initialise vectors to keep track of
summed x- and y-positions for each group (necessary for calculations later on)

sumfreqs_y = matrix(0,nrow=chains,ncol=n)

# Fill in frequency matrix and vectors using starting group configuration

for (chain in 1:chains) {

        for (i in 1:n) {

                frequencies[chain,i] = sum(group[chain,]==i)

                sumfreqs_x[chain,i] = sum(datax[group[chain,]==i])

                sumfreqs_y[chain,i] = sum(datay[group[chain,]==i])

                }

        }

# Initialise objects to store surface and convergence measures

#zmat = array(0,dim=c(chains,gridsize2,gridsize2))

#zmat2 = array(0,dim=c(chains,gridsize2,gridsize2))

#convergence = rep(NA,maxburnin)


#### Run burnin loop

for (i in 1:maxburnin) {


# Loop through all chains

for (chain in 1:chains) {


        # Perform Gibbs sampling on group allocation

        for (j in 1:n) {

                # Subtract group from frequency matrix and other objects
```

```
                frequencies[chain,group[chain,j]] =
frequencies[chain,group[chain,j]]-1

                sumfreqs_x[chain,group[chain,j]] = sumfreqs_x[chain,group[chain,j]]
- datax[j]

                sumfreqs_y[chain,group[chain,j]] = sumfreqs_y[chain,group[chain,j]]
- datay[j]

                # Calculate quantities relevant to likelihood calculation

                epsilon = 1/sqrt(frequencies[chain,]/sigma^2+1/tau^2)

                thetax = (sumfreqs_x[chain,]/sigma^2 + priorx/tau^2)*epsilon^2

                thetay = (sumfreqs_y[chain,]/sigma^2 + priory/tau^2)*epsilon^2

                # Calculate vector of likelihoods for each possible grouping

                probvec = log(frequencies[chain,])

                nextgroup = which(frequencies[chain,]==0)[1]

                probvec[nextgroup] =
integrated_prob[sum(frequencies[chain,]>0)+1]-
integrated_prob[sum(frequencies[chain,]>0)]

                probvec =
probvec+dnorm(datax[j],mean=thetax,sd=sqrt(sigma^2+epsilon^2),log=T)+dnorm(d
atay[j],mean=thetay,sd=sqrt(sigma^2+epsilon^2),log=T)

                probvec = exp(probvec-max(probvec))

                # Sample from probvec and update relevant objects

                newgroup = sample(n,1,prob=probvec)

                group[chain,j] = newgroup

                frequencies[chain,newgroup] = frequencies[chain,newgroup]+1

                sumfreqs_x[chain,newgroup] = sumfreqs_x[chain,newgroup] +
datax[j]

                sumfreqs_y[chain,newgroup] = sumfreqs_y[chain,newgroup] +
datay[j]

                }
```

```
        # Update surface

        #tempmat = matrix(0,nrow=gridsize2,ncol=gridsize2)

        uniques = unique(group[chain,])

        for (j in 1:length(uniques)) {

                # calculate posterior quantities based on grouping

                post_var = 1/(sum(group[chain,]==uniques[j])/sigma^2+1/tau^2)

                post_xmean = (sum(datax[group[chain,]==uniques[j]])/sigma^2 +
priorx/tau^2)*post_var

                post_ymean = (sum(datay[group[chain,]==uniques[j]])/sigma^2 +
priory/tau^2)*post_var

                # update temporary matrix

                #tempmat = tempmat +
dnorm(xmat,mean=post_xmean,sd=sqrt(post_var))*dnorm(ymat,mean=post_ymean,
sd=sqrt(post_var))

                }

        # add temporary matrix to surface (and squared matrix for variance
calculations)

        #zmat[chain,,] = zmat[chain,,] + tempmat/sum(tempmat)

        #zmat2[chain,,] = zmat2[chain,,] + (tempmat/sum(tempmat))^2


        # optional plot of MCMC grouping for this chain

        plot(datax,datay,pch=20,xlim=c(xmin,xmax),ylim=c(ymin,ymax),col=MCM
Ccols2[group[chain,]],main=paste(i,chain))


        } # End of chain loop


        #### Calculate Gelman Rubin Diagnostic on entire surface

        #W = colMeans(zmat2/i - (zmat/i)^2)

        #B = i*chains/(chains-1)*(colMeans((zmat/i)^2)-colMeans(zmat/i)^2)
```

```
        #R = sqrt(1+1/i*(B/W-1))

        # focus on maximum R (most un-converged cell of the surface)

        #convergence[i] = max(R)

        # plot GR diagnostic over time

        #plot(1:maxburnin,convergence,type="l",ylim=c(0.5,3.5),main=paste("Burni
n:",i))

        #abline(h=c(1,1.05),col=2,lty=2)

        #abline(v=minburnin,col=3)

        ## break burnin loop if convergence reached

        #if (max(R)<1.05 & i>=minburnin) {

        #       break

        #       }


        } # End of burnin loop



#### START MCMC MAIN LOOP  ----------------------------------------------------


#### Initialize (or re-initialize) certain objects

zmat = array(0,dim=c(chains,gridsize2,gridsize2))

zmat2 = array(0,dim=c(chains,gridsize2,gridsize2))

SE = rep(NA,maxiterations)   #vector of standard error over time

source_marginal = rep(0,n)    #stores marginal likelihood of groupings (for barplot)

#dist_marginal = array(0,dim=c(n,gridsize,gridsize))          #stores distributions
associated with each marginal likelihood (surface brakdown)

groupkeep = NULL


#### Run sample loop (only actions that differ from the burnin loop have been
```

```
commented)

for (i in 1:maxiterations) {


for (chain in 1:chains) {


        for (j in 1:n) {

                frequencies[chain,group[chain,j]] =
frequencies[chain,group[chain,j]]-1

                sumfreqs_x[chain,group[chain,j]] = sumfreqs_x[chain,group[chain,j]]
- datax[j]

                sumfreqs_y[chain,group[chain,j]] = sumfreqs_y[chain,group[chain,j]]
- datay[j]

                epsilon = 1/sqrt(frequencies[chain,]/sigma^2+1/tau^2)

                thetax = (sumfreqs_x[chain,]/sigma^2 + priorx/tau^2)*epsilon^2

                thetay = (sumfreqs_y[chain,]/sigma^2 + priory/tau^2)*epsilon^2

                probvec = log(frequencies[chain,])

                nextgroup = which(frequencies[chain,]==0)[1]

                probvec[nextgroup] =
integrated_prob[sum(frequencies[chain,]>0)+1]-
integrated_prob[sum(frequencies[chain,]>0)]

                probvec =
probvec+dnorm(datax[j],mean=thetax,sd=sqrt(sigma^2+epsilon^2),log=T)+dnorm(d
atay[j],mean=thetay,sd=sqrt(sigma^2+epsilon^2),log=T)

                probvec = exp(probvec-max(probvec))

                newgroup = sample(n,1,prob=probvec)

                group[chain,j] = newgroup

                frequencies[chain,newgroup] = frequencies[chain,newgroup]+1

                sumfreqs_x[chain,newgroup] = sumfreqs_x[chain,newgroup] +
datax[j]

                sumfreqs_y[chain,newgroup] = sumfreqs_y[chain,newgroup] +
```

```
datay[j]
                }

        tempmat = matrix(0,nrow=gridsize2,ncol=gridsize2)
        uniques = unique(group[chain,])
        for (j in 1:length(uniques)) {
                post_var = 1/(sum(group[chain,]==uniques[j])/sigma^2+1/tau^2)
                post_xmean = (sum(datax[group[chain,]==uniques[j]])/sigma^2 +
priorx/tau^2)*post_var
                post_ymean = (sum(datay[group[chain,]==uniques[j]])/sigma^2 +
priory/tau^2)*post_var
                tempmat = tempmat +
dnorm(xmat,mean=post_xmean,sd=sqrt(post_var))*dnorm(ymat,mean=post_ymean,
sd=sqrt(post_var))
                }
        zmat[chain,,] = zmat[chain,,] + tempmat/sum(tempmat)
        zmat2[chain,,] = zmat2[chain,,] + (tempmat/sum(tempmat))^2

        # Store marginal likelihood and marginal distributions
        source_marginal[length(uniques)] = source_marginal[length(uniques)]+1
        #dist_marginal[length(uniques),,] = dist_marginal[length(uniques),,] +
tempmat/sum(tempmat)

        }
        groupkeep = rbind(groupkeep,group)

        # Calculate maximum standard error of any cell in the final surface
        SE[i] = max(sqrt(colMeans(zmat2)/i-(colMeans(zmat)/i)^2)/sqrt(i))
        # plot the standard error over time
```

```
        plot(1:maxiterations,SE,type="l",ylim=c(0,10*maxSE),main=paste("Sample:
",i))
        abline(h=maxSE,col=2,lty=2)
        abline(v=miniterations,col=3)
        # break sample loop if enough samples obtained
        if (SE[i]<maxSE & i>=miniterations) {
                break
                }

        }

#### Geoprofile is final surface (normalised)

Geoprofile = colMeans(zmat)/sum(colMeans(zmat))

#### Ordermat is the final matrix of hit scores

ordermat = matrix(0,gridsize2,gridsize2)
profile_order = order(Geoprofile)
for (i in 1:gridsize2^2) {
        ordermat[profile_order[i]] = i
        }
ordermat2 = ordermat
ordermat2[ordermat2<0.95*gridsize2^2] = 0.95*gridsize2^2
hitscoremat = 1-ordermat/gridsize2^2
hitscoremat2 = 1-ordermat2/gridsize2^2
```

```
#### Create coassignment matrix for threshold grouping


coassign = matrix(0,n,n)
for (i in 1:n) {
        for (j in 1:n) {
                coassign[i,j] = mean(groupkeep[,i]==groupkeep[,j])
                }
        }
thresholdgroups = rep(8,n)
for (i in 1:n) {
        z = (coassign[i,]>(0.9))
        if (sum(z)>1) {
                thresholdgroups[z]=i
                }
        }


#### PLOT RESULTS  ------------------------------------------------------------


#Posterior groups histogram and threshold grouping
par(mfrow=c(1,2))
barplot(source_marginal[min(which(source_marginal!=0))
:max(which(source_marginal!=0))
],space=0,names.arg=min(which(source_marginal!=0))
:max(which(source_marginal!=0))
,xlab="No. Sources", ylab="Likelihood",main="Source Marginal Likelihood")
plot(datax,datay,col=MCMCcols2[thresholdgroups],pch=20,xlim=c(xmin,xmax),yli
m=c(ymin,ymax),xlab="Longitude",ylab="Latitude",main="Threshold Grouping")
```

```
#Google map

par(mfrow=c(1,1))

close.screen(all=T)

fig.mat<-matrix(c(0.0,0.7,0.0,1,0.7,1,0.0,1),nrow=2,byrow=T)

split.screen(fig.mat)

screen(1)

levels=seq(0,1,length.out=nring+1)

#DefaultMap(MyMap,xvec=xvec,yvec=yvec,Surface=1-
hitscoremat,levels=levels,transp=0.4,data=data,msources=msources)

ZoomMap(MyMap,xvec=xvec,yvec=yvec,Surface=1-
hitscoremat,levels=levels,transp=0.4,data=data,msources=msources,maptype=MapT
ype)

#points(othersources[,1],othersources[,2],pch=15,col=4)

#points(msources[,1],msources[,2],pch=15,col=2)

screen(2)
```

```
par(mar=c(0.1,0,0,0))

plot(1:10,1:10, type="n", axes=F, xlab="", ylab="")

leg.text<- c("Sigma = ", "Tau = ")

legend(0.5,9, paste(leg.text,formatC(c(sigma,tau))),bty="n",cex=c(0.6))

legend(1,7.5,c("Cases","Sources"),pch=c(20,15),bty="n",cex=c(0.6),col=c(1,4))


lengthbox<-4/(nring+1)


for (i in 1:(nring+1)) {


polygon(c(1,1,3,3),c(6-(lengthbox*i),6-(lengthbox*(i+1)),6-(lengthbox*(i+1)),6-
(lengthbox*i)),col=heat.colors(nring)[i],cex=0.6)


}


text(3.3,6.2,"Hit Score Percentage",cex=0.6)


levels2<- rev(levels)


for (i in 1:(nring+1)) {


text(4,(6-(lengthbox*i)),levels2[i],cex=0.6)


}
```

```
close.screen(all=T)




#### SAVE OPTIONS


filename = "DiagnosticPlots.png"

filename = "GPimage.png"

png(paste(Location,filename,sep=""),width=640,height=640)

png(paste(Location,filename,sep=""),width=1280,height=640)

#evaluate plot

dev.off()


#### REPORT HITSCORES?


if (length(msources)>0) {

        xdiff = abs(outer(rep(1,nrow(msources)),xvec)-
outer(msources[,1],rep(1,gridsize2)))

        ydiff = abs(outer(rep(1,nrow(msources)),yvec)-
outer(msources[,2],rep(1,gridsize2)))


        msourcex = mapply(which.min,x=split(xdiff,row(xdiff)))

        msourcey = gridsize2-(mapply(which.min,x=split(ydiff,row(ydiff))))+1


        if (nrow(msources)>1) {
```

```
            hitscores = diag(hitscoremat[msourcey,msourcex])
        } else {
            hitscores = hitscoremat[msourcey,msourcex]
        }
        hit_output = cbind(msources,hitscores)
        print(hit_output)
}



#persp(1-hitscoremat, theta = 40, phi =
15,d=5,shade=0.6,expand=0.5,box=T,border=F,xlab="long",ylab="lat",zlab="p")

#persp(Geoprofile, theta = 40, phi =
15,d=5,shade=0.6,expand=0.5,box=T,border=F,xlab="long",ylab="lat",zlab="p")
```

Appendix C: Dispersal data

Table A.1 Common.name = one or more common names, NA = not available, Type= animal (insect, mollusc, mammal, bird, fish etc), plant (conifer, herbaceous, tree, etc), fungi, Dispersal = wind, water, flight, animal vector, human etc. Full series: Y: all data available from paper; N: point estimates only; U data extracted from maps or histograms.

Common Name	Species	Type	Dispersal	Median dispersal (km)	Mean dispersal (km)	Max dispersal (km)	Full series	Reference
gray wolf	*Canis lupus*	Animal (Mammal)	Foot	NA	NA	302	U	Fritts *et al.* (1984)
gray wolf	*Canis lupus*	Animal (Mammal)	Foot	NA	NA	282	N	Henshaw & Stephenson (1974)
white wormwood	*Artemisia herba-alba*	Plant (shrub)	Seed.Gravity	0.00035	0.00035	NA	Y	Friedman & Orshan (1975)
coyote	*Canis latrans*	Animal (Mammal)	Foot	NA	NA	48	U	Danner & Fisher (1977)
redbacked vole	*Clethrionomys gapperi*	Animal (Mammal)	Foot	NA	NA	0.6	U	Bovet (1980)
cougar	*Felis concolor*	Animal (Mammal)	Foot	NA	NA	494	U (probably)	Ruth, *et al.* (1998)
snowshoe hare	*Lepus americanus*	Animal (Mammal)	Foot	NA	NA	4.83	y	Keith & Waring (1956)
black-tailed jackrabbit	*Lepus californicus*	Animal (Mammal)	Foot	NA	NA	1.61	U	Lechleitner (1958)
american marten	*Martes americana*	Animal (Mammal)	Foot	NA	NA	158	U	Slough (1989)
fisher	*Martes pennanti*	Animal (Mammal)	Foot	NA	NA	163	U	Roy (1991)
meadow vole	*Microtus pennsylvanicus*	Animal (Mammal)	Foot	NA	NA	1.2	U	Ostfeld & Manson (1996)

deer mouse	*Peromyscus maniculatus*	Animal (Mammal)	Foot	NA	NA	3.22	U	Murie & Murie (1931)
common raccoon	*Procyon lotor*	Animal (Mammal)	Foot	NA	NA	23.4	U	Tabatabai & Kennedy (1989)
common raccoon	*Procyon lotor*	Animal (Mammal)	Foot	NA	NA	29.5	U	Tabatabai & Kennedy (1989)
townsend's mole	*Scapanus townsendii*	Animal (Mammal)	Foot	NA	NA	0.14	U	Giger (1973)
grey squirrel	*Sciurus carolinensis*	Animal (Mammal)	Foot	NA	NA	4.49	U	Hungerford & Wilder (1941)
brush rabbit	*Sylvilagus bachmani*	Animal (Mammal)	Foot	NA	NA	0.16	U	Chapman (1971)
eastern cottontail rabbit	*Sylvilagus floridanus*	Animal (Mammal)	Foot	NA	NA	7.65	N	Applegate (1977)
eastern cottontail rabbit	*Sylvilagus floridanus*	Animal (Mammal)	Foot	NA	NA	19.32	U	Bowers (1954)
eastern chipmunk	*Tamias striatus*	Animal (Mammal)	Foot	NA	NA	0.55	U	Seidel (1961)
red squirrel	*Tamiasciurus hudsonicus*	Animal (Mammal)	Foot	NA	NA	1.61	N	Hamilton (1939)
valley pocket gopher	*Thomomys bottae*	Animal (Mammal)	Foot	NA	0.06	0.27	Y (?)	Vaughan (1963)
northern pocket gopher	*Thomomys talpoides*	Animal (Mammal)	Foot	NA	0.24	0.79	Y (?)	Vaughan (1963)
black bear	*Ursus americanus*	Animal (Mammal)	Foot	NA	NA	99	U	Rutherglen & Herbison (1977)
black bear	*Ursus americanus*	Animal (Mammal)	Foot	NA	NA	179	U	Payne (1975)
brown bear	*Ursus arctos*	Animal (Mammal)	Foot	NA	NA	258	U	Miller & Ballard (1982)
red fox	*Vulpes vulpes*	Animal (Mammal)	Foot	NA	NA	56.35	U	Phillips, & Mech, (1970)
sea palm	*Postelsia palmaeformis*	Algae	Water	NA	0.003	NA	N	Dayton (1973)
seaweed	*Enteromorpha*	Algae	Water	NA	35	NA	U	Amsler, & Searles (1980)
seaweed	*Enteromorpha*	Algae	Water	NA	35	NA	U	Jones & Barb (1968)
seaweed	*Enteromorpha*	Algae	Water	NA	35	NA	U	Zechman & Mathieson (1985)

giant kelp	*Macrocystis pyrifera*	Algae	Water	NA	0.01 - 0.04	NA	U	Anderson & North (1966)
giant kelp	*Macrocystis pyrifera*	Algae	Water	NA	0.01 - 0.04	NA	N	Reed *et al.* (1988)
stalked kelp	*Pterygophora californica*	Algae	Water	NA	0.5	NA	N	Reed *et al.* (1988)
filamentous brown algae	*Ectocarpus siliculosus*	Algae	Water	NA	≥4	NA	N	Reed *et al.* (1988)
NA	*Colpomenia peregrina*	Algae	Water	NA	<0.003	NA	U	Vandermeulen & DeWreede (1986)
NA	*Codium fragile spp tomentosoides*	Algae	Water	NA	12	NA	U	Carlton & Scanlon (1985)
NA	*Caulerpa taxifolia*	Algae	Water	NA	0.5	NA	U	Meinesz *et al.* (1993)
NA	*Sargassum muticum*	Algae	Water	NA	<0.005	NA	N	Andrew& Viejo (1998)
NA	*Sargassum muticum*	Algae	Water	NA	<0.005	NA	U	Deysher, & Norton (1982)
NA	*Sargassum muticum*	Algae	Water	NA	28	NA	U	Espinoza (1990)
NA	*Sargassum muticum*	Algae	Water	NA	<90	NA	U	Espinoza (1990)
NA	*Sargassum muticum*	Algae	Water	NA	10─13	NA	U	Espinoza (1990)
NA	*Sargassum muticum*	Algae	Water	NA	43	NA	U	Espinoza (1990)
orange cup coral	*Balanophyllia elegans*	Coral	Water	NA	0.0001-0.0005	NA	N	Gerrodette (1981)
stony corals	*Acroporids*	Coral	Water	NA	≤0.6	NA	N	Sammarco & Andrews (1989)
cauliflower coral	*Pocilloporids*	Coral	Water	NA	≤0.6	NA	N	Sammarco & Andrews (1989)
urn ascidian	*Didemnum molle*	Coral	Water	NA	<0.050	NA	N	Olson (1983)
european hare	*Lepus europaeus*	Animal (Mammal)	Foot	NA	NA	1.0-3.0	U	Pielowski (1972)
NA	*Diplosoma similis*	Coral	Water	NA	0.0022±0.0018	NA	Y	Stoner (1990)
european hare	*Lepus europaeus*	Animal (Mammal)	Foot	NA	NA	3.15	U	Douglas (1970)
sea squirt	*Lissoclinum patella*	Coral	Water	NA	<0.010	Y	N	Olson & McPherson (1987)
blue bell tunicate	*Podoclavella moluccensis*	Coral	Water	NA	0.0022	0.00005-0.0134	Y (?)	Davis & Butler (1989)
chain sea squirts	*Botrylloides sp*	Coral	Water	0.0006	NA	NA	Y	Worcester (1994)
chain sea squirts	*Botrylloides sp*	Coral	Water	0.225	NA	NA	N (?)	Worcester (1994)
star ascidian	*Botryllus schlosseri*	Coral	Water	NA	<0.001 (no data)	NA	N (?)	Grosberg (1987)

brown bryozoan	*Bugula neritina*	Bryozoan	Water	NA	<0.1 (where from?)	NA	N	Keough & Chernoff (1987)
giant triton	*Cymatium parthenopeum*	Mollusk	Water	NA	4400 (where from?)	NA	N	Scheltema (1971)
common periwinkle	*Littorina littorea*	Mollusk	Water	NA	42±40	NA	U	Bequaert (1943)
european hare	*Lepus europaeus*	Animal (Mammal)	Foot	1.615	NA	17.35	Y	Bray *et al*. (2007)
common periwinkle	*Littorina littorea*	Mollusk	Water	NA	42±40	NA	U	Thorson (1946)
common periwinkle	*Littorina littorea*	Mollusk	Water	NA	42±40	NA	U	Vermeij (1978)
blacklip abalone	*Haliotis rubra*	Mollusk	Water	NA	<0.015 (no data)	NA	N	Prince *et al.* (1987)
razor clam	*Ensis directus*	Mollusk	Water	NA	111	NA	U	Kenchington *et al.* (1998)
brown mussel	*Perna perna*	Mollusk	Human	NA	235	>1300	U	Hicks, & Tunnell (1995)
acorn barnacle	*Elminius modestus*	Crustacean	Animal/Human	NA	41±33	64.37-80.47	N	Crisp (1958)
spotless anemone snapping shrimp	*Alpheus immaculatus*	Crustacean		NA	0.03	NA	N	Knowlton & Keller (1986)
Japanese shore crab	*Hemigrapsus penicillatus*	Crustacean	Human	NA	160	NA	N	Noel *et al.* (1997)
asian shore crab	*Hemigrapsus sanguineus*	Crustacean	Water/Human	NA	33	NA	U	McDermott (1998)
common shore crab	*Carcinus maenas*	Crustacean	Human/Rafting	NA	173±161	NA	U	Chew (1998)
common shore crab	*Carcinus maenas*	Crustacean	Human/Rafting	NA	63	NA	U	Chew (1998)
bluestripe snapper	*Lutjanus kasmira*	Fish	Water	NA	33-130	NA	N	Chew (1998)
bluestripe snapper	*Lutjanus kasmira*	Fish	Water	NA	33-130 (no from the article)	NA	N	Randall (1987)
bluestripe snapper	*Lutjanus kasmira*	Fish	Water	NA	33-130 (no from the article)	NA	N	Randall *et al.* (1993)
tidepool sculpin	*Oligocottus maculosus*	Fish	Water	NA	<1	NA	U	Marliave (1986)
flammulated owl	*Otus flammeolus*	Animal (Mammal)	Foot	0.505	NA	0.320-0.845	N	Linkhart & Reynolds (2007)
dwarf eelgrass	*Zostera japonica*	Plant (aquatic)	Water/Bird/Human	NA	6	NA	U	Harrison & Bigley (1982)
northern spotted owl	*Strix occidentalis caurina*	Animal (Bird)	Flight	14.6	NA	0.6-111.2	Y	Forsman *et al.* (2002)
northern spotted owl	*Strix occidentalis caurina*	Animal (Bird)	Flight	24.5	NA		Y	Forsman *et al.* (2002)
northern spotted owl	*Strix occidentalis caurina*	Animal (Bird)	Flight	13.5	NA	1.8-103.5	Y	Forsman *et al.* (2002)

northern spotted owl	*Strix occidentalis caurina*	Animal (Bird)	Flight	22.9	NA		Y	Forsman *et al.* (2002)
northern spotted owl	*Strix occidentalis caurina*	Animal (Bird)	Flight	3.5	NA	NA	U	Forsman *et al.* (2002)
valley pocket gopher	*Thomomys bottae*	Animal (Mammal)	Foot	NA	0.06	NA	N	Vaughan (1963)
northern pocket gopher	*Thomomys talpoides*	Animal (Mammal)	Foot	NA	0.17	NA	N	Vaughan (1963)
NA	*Euphorbia boetica*	Plant (perennial)	Wind/Animal	0.00156	NA	0.008	Y	Narbona *et al.* (2005)
NA	*Euphorbia nicaeensis*	Plant (perennial)	Wind/Animal	0.00132	NA	0.005	Y	Narbona *et al.* (2005)
curvenut combseed	*Pectocarya recurvata*	Plant (herb)	Wind	NA	0.0007	NA	N	Venable *et al.* (2008)
curvenut combseed	*Pectocarya recurvata*	Plant (herb)	Wind	NA	0.00148	NA	N	Venable *et al.* (2008)
common mediterranean grass	*Schismus barbatus*	Plant (grass)	Wind	NA	0.00029	NA	N	Venable *et al.* (2008)
common mediterranean grass	*Schismus barbatus*	Plant (grass)	Wind	NA	0.00043	NA	N	Venable *et al.* (2008)
coconut	*Cocos spp*	Plant (Tree)	Water	NA	NA	>100	N	Ward & Brookfield. (1992)
common eelgrass	*Zostera marina L*	Plant (Seagrass)	Water	NA	NA	108.6	N	Harwell *et al.* (2002)
common eelgrass	*Zostera marina L*	Plant (Seagrass)	Water	NA	NA	34	N	Harwell *et al.* (2002)
grey mangrove	*Avicennia marina*	Plant (Tree)	Water	NA	NA	50	Y	Narbona *et al.* (2005)
turtle grass	*Thalassia testudinum*	Plant (Seagrass)	Water	NA	NA	15	N	Kaldy & Dunton (1999)
twoneedle pinyon	*Pinus edulis*	Plant (Tree)	Bird	NA		22	Y	Cain *et al.* (1998)
fireweed	*Epilobium angustifolium L*	Plant (Herb)	Wind	NA		10	Y	Cain *et al.* (1998)
devil's horsewhip	*Achyranthes aspera L*	Plant (Herb)	Adhesive	NA		4.4	Y	Cain *et al.* (1998)
black kite	*Milvus migrans*	Animal (Bird)	Flight	4.78	NA	33	Y	Forero *et al.*(2002)
northern spotted owl	*Strix occidentalis caurina*	Animal (Bird)	Flight	3.5	NA	NA	U	Forsman *et al.* (2002)
spotted owl	*Strix occidentalis*	Animal (Bird)	Flight	7	NA	NA	Y	Blakesley *et al.* (2006)
burrowing owl	*Athene cunicularia*	Animal (Bird)	Flight	NA	NA	1,860	N	Holroyd *et al.* (2011)
cougar	*Puma concolor*	Animal (Mammal)	Foot	NA	NA	483	Y	Logan, & Sweanor (2000)

cougar	*Puma concolor*	Animal (Mammal)	Foot	NA	NA	214.9	N	Sweanor *et al.* (2000)
swift fox	*Vulpes velox*	Animal (Mammal)	Foot	NA	NA	32	U	Sovada *et al.* (2003)
swift fox	*Vulpes velox*	Animal (Mammal)	Foot	NA	NA	63	N	Sovada *et al.* (2003)
swift fox	*Vulpes velox*	Animal (Mammal)	Foot	NA	NA	67	N	Sovada *et al.* (2003)
swift fox	*Vulpes velox*	Animal (Mammal)	Foot	NA	2.5-63.5	NA	?	Sovada *et al.* (2003)
swift fox	*Vulpes velox*	Animal (Mammal)	Foot	NA	2.1-61.0	NA	?	Sovada *et al.* (2003)
swift fox	*Vulpes velox*	Animal (Mammal)	Foot	NA	7.7-67.7	NA	?	Sovada *et al.* (2003)
beaver	*Castor canadensis*	Animal (Mammal)	Foot	NA	15.7	NA	Y	Beer (1955)
beaver	*Castor canadensis*	Animal (Mammal)	Foot	NA	3.7	NA	U	Beer (1955)
common blackbird	*Turdus merula*	Animal (bird)	Flight	Y	Y	Y	Y	Paradis *et al.* (1998)
white tailed deer	*Odocoileus virginianus*	Animal (Mammal)	Foot	Y	Y	Y	Y	Diefenach *et al.* (2008)
Wooly Monkey	*(Lagothrix lagothricha*	Animal (Mammal)	Brachiation	Y	Y	Y	Y	Stevenson (2000)
Canada wild ginger	*Asarum canadense*	Plant (herb)	Ant	Y	Y	Y	Y	Cain (1998)
smooth rockcress	*Arabis laevigata*	Plant (herb)	Wind	Y	Y	Y	Y	Bloom *et al.* (2002)
northern goshawk	*Accipiter gentilis*	Animal (Bird)	Flight	Y	Y	Y	Y	Weins *et al.*(2006)
savannah sparrow	*Passerculus sandwichensis*	Animal (Bird)	Flight	Y	y	Y	Y	Wheelwright & Mauck (1998)
Wild boar	*Sus scrofa*	Animal (Mammal)	Foot	Y	Y	Y	Y	Truve & Lemel (2003)
Lesser Kesterel	*Falco naumanni*	Anima (Bird)	Flight	Y	Y	Y	Y	Serrano *et al.* (2003)
NA	*Anelosimus jucundus*	Invertebrate (spider)	Foot	Y	Y	Y	Y	Avlles & Gelsey (1998)
Great Bustard	*Otis tarda*	Animal (Bird)	Flight	Y	Y	Y	Y	Alonso *et al.* (1998)
ortolan bunting	*Emberzia hortulana*	Anima (Bird)	Flight	Y	Y	Y	Y	Dale *et al.* (2004)
flying squirrel	*pteromys volans*	Animal (Mammal)	Gliding/Foot	Y	Y	Y	Y	Hanski & Selonen (2009)

white crowned sparrow	*zonotricha leucophrys*	Anima (Bird)	Flight	Y	Y	Y	Y	Morton (1997)
Black grouse	*Tetrao tetrix*	Anima (Bird)	Flight	Y	Y	Y	Y	Warren & Baines (2002)
Grey sided vole	*Clethrionomys rufocamus*	Anima (Mammal)	Foot	Y	Y	Y	Y	Saitoh (1995)
Subsoical spider	*Anelosimus cf jucundus*	Invertebrate (spider)	Foot	Y	Y	Y	Y	Powers & Aviles (2003)
Blue footed booby	*Sula nebouxii*	Animal (Bird)	Flight	Y	Y	Y	Y	Osorio-Beristain & Drummond (1993)
NA	*A glaia aff Flavida*	Bird/Wind	Air/Bird	Y	Y	Y	Y	Mack (1995)
Tree Swallow	*Tachycineta bicolor*	Animal (Bird)	Flight	Y	Y	Y	Y	Winkler *et al.* (2005)
Screech Owl	*Otus asio*	Animal (Bird)	Flight	Y	Y	Y	Y	Belthoff & Ritchison (1989)
Starling	*Sturnus vulgaris*	Animal (Bird)	Flight	Y	Y	Y	Y	Caswell *et al.* (2003)
Pied Flycatcher	*Ficedula hypoleuca*	Animal (Bird)	Flight	Y	Y	Y	Y	Caswell *et al.* (2003)
swift fox	*Vulpes velox*	Anima (Mammal)	Foot	Y	Y	Y	Y	Nicholson *et al*. (2007)
Lynx	*Lynx lynx*	Anima (Mammal)	Foot	Y	Y	Y	Y	Zimmermann *et al*. (2005)
NA	*Larrea tridentata*	Plant (herb)	Wind	NA	NA	NA	Y	Venable *et al.* (2008)
NA	*Ambrosia deltoidea*	Plant (herb)	Wind	NA	NA	NA	Y	Venable *et al.* (2008)

Appendix D: R script for the new DPM model

```
####    GP_DirichletProcess_RgoogleMaps.R

#       Description:

#       Similar to GP_DirichletProcess2.R, except able to fit Sigma using an Inverse Gamma
prior

#       Produces Googlemaps output

#       Author: Robert Verity, Mark Stevenson

#       Date: 13/07/2013

# DECLARE FUNCTIONS AND INSTALL NECCESSARY PACKAGES  #

install.packages("RgoogleMaps")

install.packages("rgdal")

install.packages("ggplot2")

install.packages("grDevices")

install.packages("coda")


library(RgoogleMaps)

library(rgdal)

library(ggplot2)

library(grDevices)
```

```
library(coda)


#### Get map based on data OR specified zoom window

GetMapClever <-
function(datax=NULL,datay=NULL,WindowLong=NULL,WindowLat=NULL,gridsize=640,maptype
="roadmap",destfile=destfile) {

        # Set zoom by data

        if (length(WindowLong)==0) {

                zoom =
max(MaxZoom(c(min(datay),max(datay)),c(min(datax),max(datax)),size=c(gridsize,gridsize)))

                center = c((min(datay)+max(datay))/2,(min(datax)+max(datax))/2)

                MyMap =
GetMap(center=center,size=c(gridsize,gridsize),zoom=zoom,maptype=maptype,destfile=destfil
e)

                }

        # Set zoom by window

        if (length(WindowLong)>0) {

                zoom = max(MaxZoom(WindowLat,WindowLong,size=c(gridsize,gridsize)))

                center = c(sum(WindowLat)/2,sum(WindowLong)/2)

                MyMap =
GetMap(center=center,size=c(gridsize,gridsize),zoom=zoom,maptype=maptype,destfile=destfil
e)

                }

        return(MyMap)

        }


#### Plot map based on saved map dimensions OR specified zoom window

PlotMap <- function(MyMap,WindowLong=NULL,WindowLat=NULL) {
```

```
	# Set zoom by data
	if (length(WindowLong)==0) {
		xwindow=c(MyMap[5]$BBOX$ll[2],MyMap[5]$BBOX$ur[2])
		ywindow=c(MyMap[5]$BBOX$ll[1],MyMap[5]$BBOX$ur[1])

	plot(1,type="n",xlim=xwindow,ylim=ywindow,xaxs="i",yaxs="i",xlab="Longitude",ylab=
"Latitude")
		rasterImage(MyMap$myTile,xwindow[1],ywindow[1],xwindow[2],ywindow[2])
		}
	# Set zoom by window
	if (length(WindowLong)>0) {

	plot(1,type="n",xlim=WindowLong,ylim=WindowLat,xaxs="i",yaxs="i",xlab="Longitude
",ylab="Latitude")

	rasterImage(MyMap$myTile,MyMap[5]$BBOX$ll[2],MyMap[5]$BBOX$ll[1],MyMap[5]$
BBOX$ur[2],MyMap[5]$BBOX$ur[1])
		}
	}

#### Convert matrix to translucent raster image
CreateRaster <- function(matrix,levels,transp) {
	tempmat = matrix(1,nrow(matrix),ncol(matrix))
	for (i in 1:(length(levels))) {
		tempmat = tempmat + ((matrix/max(matrix))>levels[i])
		}
	colvec <<- c("transparent",heat.colors(length(levels)-1))
	transp.bit = round(transp*255)
```

```
	transp.string = as.hexmode(transp.bit)
	if (transp.bit<16) {transp.string=paste("0",transp.string,sep="")}
	for (i in 2:length(colvec)) {
		colvec[i] = paste(substr(colvec[i],1,7),transp.string,sep="")
		}
	outmat = matrix(colvec[tempmat],nrow=nrow(matrix))
	return(outmat)
	}


#### Display contours that work for any zoom level
Contours <- function(xvec,yvec,matrix,levels) {
	flipmat = t(matrix[nrow(matrix):1,])
	conts = contourLines(xvec,yvec,flipmat,levels=levels)
	for (i in 1:length(conts)) {
		lines(conts[[i]]$x,conts[[i]]$y,col="dark grey")
		}
	}


#### Plot default Google map with surface and points
DefaultMap <-
function(MyMap,xvec=NULL,yvec=NULL,Surface=NULL,levels=NULL,transp=NULL,data=NULL,m
sources=NULL) {
	PlotMap(MyMap)
	if (length(Surface)>0 & length(levels)>0 & length(transp)>0 & length(xvec)>0 &
length(yvec)>0) {

	rasterImage(CreateRaster(Surface,levels,transp),MyMap$BBOX$ll[2],MyMap$BBOX$ll[
```

```
1],MyMap$BBOX$ur[2],MyMap$BBOX$ur[1])
                Contours(xvec,yvec,Surface,levels)
                }
        if (length(data)>0) {
                points(data[,1],data[,2],pch=20,cex=0.8)
                }
        if (length(msources)>0) {
                points(msources[,1],msources[,2],pch=15,col=4)
                }
        }

#### As above with zoom capability

ZoomMap <-
function(MyMap,xvec=NULL,yvec=NULL,Surface=NULL,levels=NULL,transp=NULL,data=NULL,m
sources=NULL,maptype=MapType) {
        DefaultMap(MyMap,xvec=xvec,yvec=yvec,Surface=Surface,levels=levels,transp=0.4,dat
a=data,msources=msources)
        ChooseWindow = locator(2)
        MyMap2 =
GetMapClever(WindowLong=sort(ChooseWindow$x),WindowLat=sort(ChooseWindow$y),dest
file=paste(Location,"test.png",sep=""),maptype=maptype)
        PlotMap(MyMap2,WindowLong=sort(ChooseWindow$x),WindowLat=sort(ChooseWin
dow$y))
        if (length(Surface)>0 & length(levels)>0 & length(transp)>0 & length(xvec)>0 &
length(yvec)>0) {

        rasterImage(CreateRaster(Surface,levels,transp),MyMap$BBOX$ll[2],MyMap$BBOX$ll[
1],MyMap$BBOX$ur[2],MyMap$BBOX$ur[1])
                Contours(xvec,yvec,Surface,levels)
```

```
		}
	if (length(data)>0) {
		points(data[,1],data[,2],pch=20,cex=0.8)
		}
	if (length(msources)>0) {
		points(msources[,1],msources[,2],pch=15,col=4)
		}
	}


#### Compute pairwise distances between data
Pairwise <- function(data) {
	xmat1 = outer(rep(1,n),data[,1])
	xmat2 = outer(data[,1],rep(1,n))
	xdist = abs(xmat1-xmat2)
	ymat1 = outer(rep(1,n),data[,2])
	ymat2 = outer(data[,2],rep(1,n))
	ydist = abs(ymat1-ymat2)
	zdist = sqrt(xdist^2+ydist^2)
	output = zdist[col(xmat1)>row(xmat1)]
	return(output)
	}


#### Import data from .txt or .csv adaptively
ImportData <- function(header=F) {
	filepath = tryCatch(file.choose(), error = function(e) NULL)
```

```
	if (length(filepath)==0) {
		cat("Import cancelled by user\n")
		output = NULL
	} else {
		extension = tail(unlist(strsplit(filepath,"[.]")),1)
		if (extension=="txt") {output = as.matrix(read.table(filepath,header=header))}
		if (extension=="csv") {output = as.matrix(read.csv(filepath,header=header))}
		return(output)
	}
}

#### density function of inverse-gamma distribution
dinvgamma <- function(x,shape,rate,log=F) {
	output = shape*log(rate)-lgamma(shape)-(shape+1)*log(x)-rate/x
	if (log==F) output = exp(output)
	return(output)
}

#### density function of the square root of an inverse-gamma distributed random variable
drootinvgamma <- function(x,shape,rate,log=F) {
	output = log(2)+shape*log(rate)-lgamma(shape)-(2*shape+1)*log(x)-rate/x^2
	if (log==F) output = exp(output)
	return(output)
}
```

```
# INPUT STARTING PARAMETERS ----------------------------------------------------


#### Import data

#data = ImportData(header=F)

#data = cbind(rnorm(60,mean=rep(c(-0.094825,-0.042622,-
0.136344),times=c(15,20,25)),sd=0.01),rnorm(60,mean=rep(c(51.491511,51.524606,51.521795
),times=c(15,20,25)),sd=0.01))

datax = data[,1]

datay = data[,2]

n = length(datax)




#### Setup the colourspace

MCMCcols <-
colorRampPalette(c('red','green','orange','blue','yellow','gray','black','brown','aquamarine3','cy
an','darkmagenta','darkviolet','green4'))

MCMCcols2 = sample(MCMCcols(n))


#### Import sources (optional)

input <- "NA"

while(!isTRUE(input=="Y") && !isTRUE(input=="N")) {

        cat("Import source data? Enter Y=Yes or N=No\n")

        input <- scan("",what="character",n=1,quiet=T)

        if (input=="Y") {

                msources = ImportData(header=F)

                if (length(msources)>0) cat("Source data imported\n")

        } else if (input=="N") {
```

```
                cat("Not importing source data\n")
                msources = NULL
        } else {cat("Incorrect input: ")}
}

#### Input simulated source locations
#msources = cbind(c(-0.094825,-0.042622,-0.136344),c(51.491511,51.524606,51.521795))

#msources=read.table(file.choose())
#othersources=read.table(file.choose())

#### Histogram pairwise distances between data
par(mfrow=c(1,1))
pairwise = Pairwise(data)
hist(pairwise,breaks=2*n,col=8)
abline(v=0.0175,col=2)

#### Visualisation parameters

nring=20                    #Number of levels for contour plots
transp=0.4                  #Transparncy of profile overlay
gridsize =  640             #Number of cells in map grid (same in both dimensions) up to 640 max
gridsize2 = 300             #Model resolution
MapType= "roadmap"   #Map type can be any one of several types "roadmap","mobile","satellite", "terrain","hybrid" etc
```

```
#Location= "C:\\Documents and Settings\\Mark Stevenson\\My Documents\\Dropbox\\Work Shared\\Bob\\"

#Location= "~/Desktop/Dropbox/Work Shared/Bob/"

Location= "~/Desktop/GP output/"

Location = "C:\\Users\\Bob\\Desktop\\"



#### Input model and MCMC parameters



minburnin = 200                    #run burnin for at least this long

maxburnin = 1000                   #run burnin for at most this long

chains = 5                         #number of chains to run simultaneously

miniterations = 100            #take samples for at least this many iterations

maxiterations = 10000          #take samples for at most this many iterations

maxSE = 0.01                   #stop taking samples when this standard error is reached (and miniterations exceeded)



# PLOT PRIOR ON MAP ------------------------------------------------



#### Download map and extract some useful measures



MyMap = GetMapClever(datax=datax,datay=datay,destfile=paste(Location,"RawMap.png",sep=""),maptype=MapType)



#MyMap =
```

```
GetMapClever(WindowLong=c(113.8,114.6),WindowLat=c(22.2,23),destfile=paste(Location,"RawMap.png",sep=""),maptype=MapType)


xmin = MyMap[5]$BBOX$ll[2]

xmax = MyMap[5]$BBOX$ur[2]

ymin = MyMap[5]$BBOX$ll[1]

ymax = MyMap[5]$BBOX$ur[1]



### Set paramteres for inverse gamma prior on sigma


sigma_expectation = (max(c((ymax-ymin),(xmax-xmin)))*0.01)

sigma_expectation = 0.02

delta = 1                    #shape parameter of inverse-gamma prior on variance

beta = sigma_expectation^2/pi            #rate parameter of inverse-gamma prior on variance

tau = "DEFAULT"                                    #standard deviation of prior on source location (Normal distribution). Set as "DEFAULT" for default


sdvec = seq(0,0.05,length.out=1001)

sdprior = drootinvgamma(sdvec,shape=delta,rate=beta)

plot(sdvec,sdprior,type="l",xlab="Standard Deviation (Longitude)",ylab="Probability")

abline(v=sigma_expectation,lty=2)


#### Create prior


priorx = (xmin+xmax)/2
```

```
priory = (ymin+ymax)/2


if (tau=="DEFAULT") {
        xdiff = max(datax)-min(datax)
        ydiff = max(datay)-min(datay)
        tau = max(c(xdiff,ydiff))
}


xvec = seq(xmin,xmax,length.out=gridsize2)
yvec = seq(ymin,ymax,length.out=gridsize2)
xmat = outer(rep(1,gridsize2),xvec)
ymat = outer(yvec[gridsize2:1],rep(1,gridsize2))
priormat = dnorm(xmat,mean=priorx,sd=tau)*dnorm(ymat,mean=priory,sd=tau)


#### Plot Google map and overlay prior
levels=seq(0,1,length.out=nring+1)
PlotMap(MyMap)
priorraster = CreateRaster(priormat,levels,transp)
rasterImage(priorraster,xmin,ymin,xmax,ymax)
par(new=T);
contour(xvec,yvec,t(priormat/max(priormat)),xaxs="i",yaxs="i",levels=levels,axes=F,drawlabels
=F,col="dark grey")
points(datax,datay,pch=20,cex=0.8,col="red")


# INTEGRATION ----------------------------------------------------
```

```
#### Integrate over hyper-prior on alpha (some fancy integration tricks to make this possible).
Output in log space, where the ith element of the vector integrated_prob contains the logged
integral of (x^i)*gamma(x)/gamma(n+x) over the hyperprior 1/(1+x)^2

integrated_prob = rep(0,n)

for (i in 1:n) {

        temp = rep(0,1001)

        for (j in 2:1001) {

                integrand = function(x) {

                        exp((j-501)*log(10) + i*log(x*n) + lgamma(x*n)-lgamma(n+x*n) -
2*log(1+x*n))

                        }

                temp[j] = integrate(integrand,lower=0,upper=Inf)$value*n

                if (temp[j]<0) temp[j]=0

                temp[j] = log(temp[j]) - (j-501)*log(10)

                if (temp[j]!=-Inf & abs(temp[j]-temp[j-1])<0.0001) {

                        integrated_prob[i] = temp[j]

                        break()

                        }

                }

        }


#### START MCMC BURNIN LOOP  ----------------------------------------------------


# Initialise Gibbs sampling objects

mux_burnin = list()

muy_burnin = list()

group_burnin = list()
```

```
frequencies_burnin = list()

sumdatax_burnin = list()

sumdatay_burnin = list()

sigma_burnin = list()

convergence = list()

for (chain in 1:chains) {

        mux_burnin[[chain]] = matrix(0,nrow=maxburnin,ncol=n)

                mux_burnin[[chain]][1,] = rnorm(n,sd=100)

        muy_burnin[[chain]] = matrix(0,nrow=maxburnin,ncol=n)

                muy_burnin[[chain]][1,] = rnorm(n,sd=100)

        group_burnin[[chain]] = matrix(1,nrow=maxburnin,ncol=n)

        frequencies_burnin[[chain]] = matrix(0,nrow=maxburnin,ncol=n)

                frequencies_burnin[[chain]][1,1] = n

        sumdatax_burnin[[chain]] = matrix(0,nrow=maxburnin,ncol=n)

                sumdatax_burnin[[chain]][1,1] = sum(datax)

        sumdatay_burnin[[chain]] = matrix(0,nrow=maxburnin,ncol=n)

                sumdatay_burnin[[chain]][1,1] = sum(datay)

        sigma_burnin[[chain]] = rep(1,maxburnin)

        convergence[[chain]] = 1

}


#### Run burnin loop

for (i in 2:maxburnin) {


# Loop through all chains
```

```
for (chain in 1:chains) {


        # Update objects with values from last iteration

        mux_burnin[[chain]][i,] = mux_burnin[[chain]][i-1,]

        muy_burnin[[chain]][i,] = muy_burnin[[chain]][i-1,]

        group_burnin[[chain]][i,] = group_burnin[[chain]][i-1,]

        frequencies_burnin[[chain]][i,] = frequencies_burnin[[chain]][i-1,]

        sumdatax_burnin[[chain]][i,] = sumdatax_burnin[[chain]][i-1,]

        sumdatay_burnin[[chain]][i,] = sumdatay_burnin[[chain]][i-1,]

        sigma_burnin[[chain]][i] = sigma_burnin[[chain]][i-1]


        # Perform Gibbs sampling on group allocation

        for (j in 1:n) {

                # Subtract this observation from frequency matrix and other objects

                frequencies_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
frequencies_burnin[[chain]][i,group_burnin[[chain]][i,j]] - 1

                sumdatax_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
sumdatax_burnin[[chain]][i,group_burnin[[chain]][i,j]] - datax[j]

                sumdatay_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
sumdatay_burnin[[chain]][i,group_burnin[[chain]][i,j]] - datay[j]

                # Draw new value of mu with this point removed

                postvar =
1/(frequencies_burnin[[chain]][i,group_burnin[[chain]][i,j]]/sigma_burnin[[chain]][i]^2+1/tau^
2)

                postmeanx =
(sumdatax_burnin[[chain]][i,group_burnin[[chain]][i,j]]/sigma_burnin[[chain]][i]^2 +
priorx/tau^2)*postvar

                postmeany =
(sumdatay_burnin[[chain]][i,group_burnin[[chain]][i,j]]/sigma_burnin[[chain]][i]^2 +
```

```
priory/tau^2)*postvar

                mux_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
rnorm(1,mean=postmeanx,sd=sqrt(postvar))

                muy_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
rnorm(1,mean=postmeany,sd=sqrt(postvar))

                # Calculate vector of likelihoods for each possible grouping

                probvec = log(frequencies_burnin[[chain]][i,])

                probvec =
probvec+dnorm(datax[j],mean=mux_burnin[[chain]][i,],sd=sigma_burnin[[chain]][i],log=T)+dno
rm(datay[j],mean=muy_burnin[[chain]][i,],sd=sigma_burnin[[chain]][i],log=T)

                nextgroup = which(frequencies_burnin[[chain]][i,]==0)[1]

                probvec[nextgroup] =
integrated_prob[sum(frequencies_burnin[[chain]][i,]>0)+1]-
integrated_prob[sum(frequencies_burnin[[chain]][i,]>0)]

                probvec[nextgroup] = probvec[nextgroup] +
dnorm(datax[j],mean=priorx,sd=sqrt(sigma_burnin[[chain]][i]^2+tau^2),log=T)+dnorm(datay[j],
mean=priory,sd=sqrt(sigma_burnin[[chain]][i]^2+tau^2),log=T)

                probvec = exp(probvec-max(probvec))    #remove underflow

                # Sample from probvec and update relevant objects

                newgroup = sample(n,1,prob=probvec)

                group_burnin[[chain]][i,j] = newgroup

                frequencies_burnin[[chain]][i,newgroup] =
frequencies_burnin[[chain]][i,newgroup] + 1

                sumdatax_burnin[[chain]][i,newgroup] =
sumdatax_burnin[[chain]][i,newgroup] + datax[j]

                sumdatay_burnin[[chain]][i,newgroup] =
sumdatay_burnin[[chain]][i,newgroup] + datay[j]

                if (newgroup==nextgroup) {

                        postvar =
1/(frequencies_burnin[[chain]][i,group_burnin[[chain]][i,j]]/sigma_burnin[[chain]][i]^2+1/tau^
2)
```

```
                    postmeanx =
(sumdatax_burnin[[chain]][i,group_burnin[[chain]][i,j]]/sigma_burnin[[chain]][i]^2 +
priorx/tau^2)*postvar

                    postmeany =
(sumdatay_burnin[[chain]][i,group_burnin[[chain]][i,j]]/sigma_burnin[[chain]][i]^2 +
priory/tau^2)*postvar

                    mux_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
rnorm(1,mean=postmeanx,sd=sqrt(postvar))

                    muy_burnin[[chain]][i,group_burnin[[chain]][i,j]] =
rnorm(1,mean=postmeany,sd=sqrt(postvar))

                }

                }


        # vary sigma conditional on grouping

        sigma_burnin[[chain]][i] = 1/sqrt(rgamma(1,shape=delta+n,rate=beta +
0.5*sum((datax - mux_burnin[[chain]][i,group_burnin[[chain]][i,]])^2) + 0.5*sum((datay -
muy_burnin[[chain]][i,group_burnin[[chain]][i,]])^2)))

        convergence[[chain]] = mcmc(c(convergence[[chain]],sigma_burnin[[chain]][i]))


        } # End of chain loop


        # optional plot of convergence

        if (floor(i/10)==(i/10) & i>50) {

                gelman.plot(convergence)

                abline(h=1.1,lty=3)

                }


        } # End of burnin loop
```

```
#### Plot trace for sigma (standard deviation) across all chains
plot(1,type="n",xlim=c(0,maxburnin),ylim=c(0,0.1),main="sigma trace")
for (chain in 1:chains) {
        points(sigma_burnin[[chain]],ylim=c(0,0.1),pch=20,cex=0.5,col=chain)
}



#### START MCMC MAIN LOOP  ----------------------------------------------------



# Initialise Gibbs sampling objects


mux = matrix(0,nrow=maxiterations,ncol=n)
        mux[1,] = mux_burnin[[1]][i,]
muy = matrix(0,nrow=maxiterations,ncol=n)
        muy[1,] = muy_burnin[[1]][i,]
group = matrix(0,nrow=maxiterations,ncol=n)
        group[1,] = group_burnin[[1]][i,]
frequencies = matrix(0,nrow=maxiterations,ncol=n)
        frequencies[1,] = frequencies_burnin[[1]][i,]
sumdatax = matrix(0,nrow=maxiterations,ncol=n)
        sumdatax[1,] = sumdatax_burnin[[1]][i,]
sumdatay = matrix(0,nrow=maxiterations,ncol=n)
        sumdatay[1,] = sumdatay_burnin[[1]][i,]
sigma = rep(0,maxiterations)
```

```
        sigma[1] = sigma_burnin[[1]][i]

sigma_rate = rep(0,maxiterations)

        sigma_rate[1] = beta + 0.5*sum((datax -
mux_burnin[[chain]][i,group_burnin[[chain]][i,]])^2) + 0.5*sum((datay -
muy_burnin[[chain]][i,group_burnin[[chain]][i,]])^2)


#### Run sample loop

for (i in 2:maxiterations) {


        # Update objects with values from last iteration

        mux[i,] = mux[i-1,]

        muy[i,] = muy[i-1,]

        group[i,] = group[i-1,]

        frequencies[i,] = frequencies[i-1,]

        sumdatax[i,] = sumdatax[i-1,]

        sumdatay[i,] = sumdatay[i-1,]

        sigma[i] = sigma[i-1]


        # Perform Gibbs sampling on group allocation

        for (j in 1:n) {

                # Subtract this observation from frequency matrix and other objects

                frequencies[i,group[i,j]] = frequencies[i,group[i,j]] - 1

                sumdatax[i,group[i,j]] = sumdatax[i,group[i,j]] - datax[j]

                sumdatay[i,group[i,j]] = sumdatay[i,group[i,j]] - datay[j]

                # Draw new value of mu with this point removed

                postvar = 1/(frequencies[i,group[i,j]]/sigma[i]^2+1/tau^2)
```

```
                postmeanx = (sumdatax[i,group[i,j]]/sigma[i]^2 + priorx/tau^2)*postvar

                postmeany = (sumdatay[i,group[i,j]]/sigma[i]^2 + priory/tau^2)*postvar

                mux[i,group[i,j]] = rnorm(1,mean=postmeanx,sd=sqrt(postvar))

                muy[i,group[i,j]] = rnorm(1,mean=postmeany,sd=sqrt(postvar))

                # Calculate vector of likelihoods for each possible grouping

                probvec = log(frequencies[i,])

                probvec =
probvec+dnorm(datax[j],mean=mux[i,],sd=sigma[i],log=T)+dnorm(datay[j],mean=muy[i,],sd=si
gma[i],log=T)

                nextgroup = which(frequencies[i,]==0)[1]

                probvec[nextgroup] = integrated_prob[sum(frequencies[i,]>0)+1]-
integrated_prob[sum(frequencies[i,]>0)]

                probvec[nextgroup] = probvec[nextgroup] +
dnorm(datax[j],mean=priorx,sd=sqrt(sigma[i]^2+tau^2),log=T)+dnorm(datay[j],mean=priory,sd
=sqrt(sigma[i]^2+tau^2),log=T)

                probvec = exp(probvec-max(probvec))    #remove underflow

                # Sample from probvec and update relevant objects

                newgroup = sample(n,1,prob=probvec)

                group[i,j] = newgroup

                frequencies[i,newgroup] = frequencies[i,newgroup] + 1

                sumdatax[i,newgroup] = sumdatax[i,newgroup] + datax[j]

                sumdatay[i,newgroup] = sumdatay[i,newgroup] + datay[j]

                if (newgroup==nextgroup) {

                        postvar = 1/(frequencies[i,group[i,j]]/sigma[i]^2+1/tau^2)

                        postmeanx = (sumdatax[i,group[i,j]]/sigma[i]^2 +
priorx/tau^2)*postvar

                        postmeany = (sumdatay[i,group[i,j]]/sigma[i]^2 +
priory/tau^2)*postvar
```

```
                    mux[i,group[i,j]] = rnorm(1,mean=postmeanx,sd=sqrt(postvar))

                    muy[i,group[i,j]] = rnorm(1,mean=postmeany,sd=sqrt(postvar))

                }

                }


        # vary sigma conditional on grouping

        sigma_rate[i] = beta + 0.5*sum((datax - mux[i,group[i,]])^2) + 0.5*sum((datay -
muy[i,group[i,]])^2)

        sigma[i] = 1/sqrt(rgamma(1,shape=delta+n,rate=beta + 0.5*sum((datax -
mux[i,group[i,]])^2) + 0.5*sum((datay - muy[i,group[i,]])^2)))


        # optional plot of MCMC grouping for this chain

        if (floor(i/100)==(i/100)) {


        plot(datax,datay,pch=20,xlim=c(xmin,xmax),ylim=c(ymin,ymax),col=MCMCcols2[group[
i,]],main=paste(i,chain))

                }


        } # End of sample loop


#### MCMC DIAGNOSTIC PLOTS  ------------------------------------------------------------


#### Four in one panel

par(mfrow=c(2,2))


# trace plot of sigma (can be removed)

plot(sigma,ylim=c(0,0.1),pch=20,cex=0.5,main="sigma trace")
```

```
# distribution of sigma

sdpost =
colMeans(mapply(drootinvgamma,x=sdvec,MoreArgs=list(shape=delta+n,rate=sigma_rate)))

plot(sdvec,sdpost,type="l",xlab="sigma",ylab="posterior density")

lines(sdvec,sdprior,lty=2)


# autocorrelation plot

autocorr.plot(sigma,auto.layout=F,lag.max=maxiterations)


#Posterior groups histogram

groupnum = rowSums(frequencies>0)

realised_sources = hist(groupnum,breaks=0:n,plot=F)$intensities

realised_sources_nonzero = min(which(realised_sources!=0)):max(which(realised_sources!=0))

barplot(realised_sources[realised_sources_nonzero],names.arg=realised_sources_nonzero,spa
ce=0,xlab="Realised Sources",ylab="Probability")



#### THINNING  ------------------------------------------------------


thinning = 500


sigma_thin = sigma[seq(1,maxiterations,thinning)]

group_thin = group[seq(1,maxiterations,thinning),]

frequencies_thin = frequencies[seq(1,maxiterations,thinning),]

sumdatax_thin = sumdatax[seq(1,maxiterations,thinning),]
```

```
sumdatay_thin = sumdatay[seq(1,maxiterations,thinning),]


#### CONSTRUCT GEOPROFILE  ------------------------------------------------------------


#### Construct geoprofile by averaging over conditional posterior distributions

postvar = 1/(frequencies_thin/outer(sigma_thin,rep(1,n))^2+1/tau^2)

postmeanx = (sumdatax_thin/outer(sigma_thin,rep(1,n))^2 + priorx/tau^2)*postvar

postmeany = (sumdatay_thin/outer(sigma_thin,rep(1,n))^2 + priory/tau^2)*postvar

Geoprofile = matrix(0,nrow=gridsize2,ncol=gridsize2)

for (i in 1:gridsize2) {

        print(paste(i,"of",gridsize2))

        flush.console()

        M1 = outer(dnorm(yvec[gridsize2+1-
i],mean=postmeany[frequencies_thin>0],sd=sqrt(postvar[frequencies_thin>0])),rep(1,gridsize2
))

        M2 =
mapply(dnorm,x=xvec,MoreArgs=list(mean=postmeanx[frequencies_thin>0],sd=sqrt(postvar[fr
equencies_thin>0])))

        Geoprofile[i,] = colSums(M1*M2)

}



#### Ordermat is the final matrix of hit scores


ordermat = matrix(0,gridsize2,gridsize2)

profile_order = order(Geoprofile)

for (i in 1:gridsize2^2) {
```

```
        ordermat[profile_order[i]] = i
        }
ordermat2 = ordermat
ordermat2[ordermat2<0.95*gridsize2^2] = 0.95*gridsize2^2
hitscoremat = 1-ordermat/gridsize2^2
hitscoremat2 = 1-ordermat2/gridsize2^2

#### Create coassignment matrix for threshold grouping

coassign = matrix(0,n,n)
for (i in 1:n) {
        for (j in 1:n) {
                coassign[i,j] = mean(group_thin[,i]==group_thin[,j])
                }
        }
thresholdgroups = rep(8,n)
for (i in 1:n) {
        z = (coassign[i,]>(0.9))
        if (sum(z)>1) {
                thresholdgroups[z]=i
                }
        }

#### PLOT RESULTS  ------------------------------------------------------------
```

```
#Google map

par(mfrow=c(1,1))

close.screen(all=T)

fig.mat<-matrix(c(0.0,0.7,0.0,1,0.7,1,0.0,1),nrow=2,byrow=T)

split.screen(fig.mat)

screen(1)

levels=seq(0,1,length.out=nring+1)

#DefaultMap(MyMap,xvec=xvec,yvec=yvec,Surface=1-
hitscoremat,levels=levels,transp=0.4,data=data,msources=msources)

ZoomMap(MyMap,xvec=xvec,yvec=yvec,Surface=1-
hitscoremat,levels=levels,transp=0.4,data=data,msources=msources,maptype=MapType)

#points(othersources[,1],othersources[,2],pch=15,col=4)

#points(msources[,1],msources[,2],pch=15,col=2)

screen(2)

par(mar=c(0.1,0,0,0))
```

```
plot(1:10,1:10, type="n", axes=F, xlab="", ylab="")

leg.text<- c("Sigma = ", "Tau = ")

#legend(0.5,9, paste(leg.text,formatC(c(sigma,tau))),bty="n",cex=c(0.6))

legend(1,7.5,c("Cases","Sources"),pch=c(20,15),bty="n",cex=c(0.6),col=c(1,4))


lengthbox<-4/(nring+1)


for (i in 1:(nring+1)) {


polygon(c(1,1,3,3),c(6-(lengthbox*i),6-(lengthbox*(i+1)),6-(lengthbox*(i+1)),6-
(lengthbox*i)),col=heat.colors(nring)[i],cex=0.6)


}


text(3.3,6.2,"Hit Score Percentage",cex=0.6)


levels2<- rev(levels)


for (i in 1:(nring+1)) {


text(4,(6-(lengthbox*i)),levels2[i],cex=0.6)


}


close.screen(all=T)
```

```
#### SAVE OPTIONS


filename = "DiagnosticPlots.png"

filename = "GPimage.png"

png(paste(Location,filename,sep=""),width=640,height=640)

png(paste(Location,filename,sep=""),width=1280,height=640)

#evaluate plot

dev.off()


#### REPORT HITSCORES?


if (length(msources)>0) {

        xdiff = abs(outer(rep(1,nrow(msources)),xvec)-outer(msources[,1],rep(1,gridsize2)))

        ydiff = abs(outer(rep(1,nrow(msources)),yvec)-outer(msources[,2],rep(1,gridsize2)))


        msourcex = mapply(which.min,x=split(xdiff,row(xdiff)))

        msourcey = gridsize2-(mapply(which.min,x=split(ydiff,row(ydiff))))+1


        if (nrow(msources)>1) {

                hitscores = diag(hitscoremat[msourcey,msourcex])

        } else {

                hitscores = hitscoremat[msourcey,msourcex]

        }
```

```
        hit_output = cbind(msources,hitscores)

        print(hit_output)

}



#persp(1-hitscoremat, theta = 40, phi =
15,d=5,shade=0.6,expand=0.5,box=T,border=F,xlab="long",ylab="lat",zlab="p")

#persp(Geoprofile, theta = 40, phi =
15,d=5,shade=0.6,expand=0.5,box=T,border=F,xlab="long",ylab="lat",zlab="p")
```

Appendix E: Alpine newt locations and verification protocol

Records of alpine newt locations and the source of the record**:**

COUNTY	FIRST RECORDED	SOURCE
Argyll	1968	Irving, D.1987. New amphibian records for Argyll. British Herpetological Society. 20,31.
Cumbria	2009	Alien Encounters via ARC trust
Cumbria	2008	Alien Encounters via ARC trust
Cumbria	"number of years"	Hesketh Ecology, Ecological Consultancy
Cumbria	2012	Hesketh Ecology, Ecological Consultancy
Tyne & Wear	2008	Countryside Officer, South Tyneside Council, via ARC trust
Tyne & Wear	1984	Banks, B. 1989. Alpine newts in north east England. British Herpetological Society Bulletin. 30, 4-5.
Tyne & Wear	pre-1981	Banks, B. 1989. Alpine newts in north east England. British Herpetological Society Bulletin. 30, 4-5.
Cleveland	2004	Bond, I. and Haycock, G. 2008. The Alpine newt in northern England. Herpetological Bulletin. 104, 4-6.
Tyne & Wear	2007	NBN Gateway (dataset North East Environmental Data Hub non-sensitive species records) via ARC trust
Tyne & Wear	2005	Francesca Leslie (EYE project officer) by email via ARC trust
Tyne & Wear	2006	Tees Valley Wildlife Trust via Ian Bond
Tyne & Wear	2006	Tees Valley Wildlife Trust via Ian Bond
Tyne & Wear	2009	Ian Bond via email
Tyne & Wear	2007	Francesca Leslie (EYE project officer) by email via ARC trust AND Tees Valley Wildlife Trust via Ian Bond
Stockton	2009	Claire Gilchrist via Ian Bond
Cleveland		2009 British Trust for Ornithology BTO Survey via ARC trust
Tyne & Wear	1998	NBN Gateway (dataset North East Environmental Data Hub non-sensitive species records) via ARC trust
Tyne & Wear	1998	Francesca Leslie (EYE project officer) by email via ARC trust
Tyne & Wear	2003	Francesca Leslie (EYE project officer) by email AND NBN Gateway (dataset North East Environmental Data Hub non-sensitive species records) via ARC trust
Tyne & Wear		ARC trust
Tyne & Wear		ARC trust
Tyne & Wear	2011	Ian Bond via email
Tyne & Wear	2011	PrismPlanning Ecological Survey
Tyne & Wear	2011	PrismPlanning Ecological Survey
Darlington	2012	Ian Bond via email
South Shields	2012	Ian Bond via email
South Shields	2009	Ian Bond via email
Yorkshire, West	2001	Bond, I. and Haycock, G. 2008. The Alpine newt in northern England. Herpetological Bulletin. 104, 4-6.
Yorkshire, West	2010-2011	Haycock & Jay Associates, Consultant Ecologists

Yorkshire, West	10/05/2011	Haycock & Jay Associates, Consultant Ecologists
Yorkshire, West	1990s	ARC trust
Yorkshire, West		ARC trust
Yorkshire, West	2001	Bond, I. and Haycock, G. 2008. The Alpine newt in northern England. Herpetological Bulletin. 104, 4-6.
Shropshire	1970	Banks, B. (1991). British newts. British Wildlife 2(6),362-365; Bell & Bell.1995. Proceedings of the 2nd World Congress of Herpetology via Arc trust
Yorkshire, South	Approx 1993	Bond, I. and Haycock, G. 2008. The Alpine newt in northern England. Herpetological Bulletin. 104, 4-6.
Staffordshire	2009	ARC trust
Bristol		2009 British Trust for Ornithology BTO Survey via ARC trust
Devon	app 2004	2009 British Trust for Ornithology BTO Survey via ARC trust
Devon	app2007	Devon Biodiversity Records Centre
Northamptonshire	WW II	Blackwell, K. 2002. *Triturus alpestris* in Britain. Herpetological Bulletin. 79,32.
Essex		ARC trust
Surrey	1920s	Beebee, T. and Griffiths, R. 2000. Amphibians and Reptiles: A Natural History of the British Herpetofauna. London: Harper Collins Publishers.
Brighton and Hove	1970s	Beebee, T. and Griffiths, R. 2000. Amphibians and Reptiles: A Natural History of the British Herpetofauna. London: Harper Collins Publishers.
Hertfordshire	app 1987	2009 British Trust for Ornithology BTO Survey via ARC trust
Essex	2009	Herpetologic Ltd
Essex	2009	Herpetologic Ltd
Essex	2009	Herpetologic Ltd
Essex	2009	Herpetologic Ltd
Kent	1980s	Sewell, D. (2006). http://www.open.ac.uk/daptf/froglog/Froglog76.pdf
East Kent	08/05/1994	NBN Gateway (dataset Reptiles and Amphibians Dataset, provided by Biological Records Centre) via ARC
Greater London		ARC trust
Edinburgh	May 2011	Friends of Angus Herpetofauna
Cornwall	24/02/2012	N. Morris, Cornwall collage, Newquay
Cornwall	24/02/2013	N. Morris, Cornwall collage, Newquay
Cornwall	24/02/2014	N. Morris, Cornwall collage, Newquay
Cornwall	24/02/2015	N. Morris, Cornwall collage, Newquay
Leicestershire	2006	Leicertershire Amphibian and reptile network
Leicestershire	2006	Leicertershire Amphibian and reptile network
Lothians		LARG (Lothian Amphibian and Reptile Group)
Lothians		LARG (Lothian Amphibian and Reptile Group)
Lothians		LARG (Lothian Amphibian and Reptile Group)
Lothians		LARG (Lothian Amphibian and Reptile Group)
Surrey	April 2012	Confidential Ecological report
Surrey	May 2012	Confidential Ecological report
Surrey	May 2012	Confidential Ecological report
Bristol	23/03/2010	Bristol Regional Environmental Records Centre
Stanmore	2007/8	Froglife

Essex	Herpetologic Ltd.
Essex	Herpetologic Ltd.
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student
Canterbury	J. Sears, PhD Student

Verification Criteria:

1. The person making the recordings' knowledge base is accepted by their peers and/or a national body.
2. The record has been deemed reliable by a trusted and experienced ecological recorder/recording organisation.
3. The locations are published in a report or ecological/planning survey.
4. There is photographic evidence.

Appendix F: The R package ‘disperse’

Package: disperse

Type: Package

Title: Transforms and plots dispersal data

Version: 1.0

Date: 2013-04-23

Author: Mark Stevenson and Robert Verity (Queen Mary University of London)

Maintainer: Mark Stevenson <m.stevenson@qmul.ac.uk>

Description: Calculates the correct transformation form two dimesnsional dispersal data to a one dimesnional histogram. Plots the resulting histogram with error bars.

LazyData: yes

License: GPL-2 | GPL-3

exportPattern("^[[:alpha:]]+")

```
fnlist <-

function(x, fil){ z <- deparse(substitute(x))

                cat(z, "\n", file=fil)

                nams=names(x)
```

```
        for (i in seq_along(x) ){ cat(nams[i], "\t",  x[[i]], "\n",

                        file=fil, append=TRUE) }

}

\name{Disperse-package}

\alias{Disperse-package}

\alias{Disperse}

\docType{package}

\title{

Transforms and plots dispersal data

~~ package title ~~

}

\description{

Dispersal is crucial in biology, with profound consequences in areas including population genetics, mate choice, metapopulation dynamics and population viability. Unfortunately, the presentation of dispersal data is marked by a simple geometrical mistake often that leads to mistaken biological inferences. Here, we compute the correct mathematical transformation from two to one dimensional dispersal as well as maximum likelohhod error bars. These are then plotted on a hsitogram.

}
```

\details{

\tabular{ll}{

Package: \tab Disperse\cr

Type: \tab Package\cr

Version: \tab 1.0\cr

Date: \tab 2013-04-23\cr

License: \GPL-2 | GPL-3\\cr

}

Transform

fnlist

exampledata

Transplot

}

\author{

M.D. Stevenson and R. Verity

Maintainer: M.D. Stevenson <m.stevenson@qmul.ac.uk>

}

\references{

Stevenson, et al (2013) Dispersal data in ecology: three golden rules. Trends in Ecology and Evolution.

}

\keyword{ package }

\seealso{

}

\examples{

EXAMPLES

}

Transform <-

function(bincounts,binwidths) {

```
# Compute some stuff


bins <<-length(bincounts)

N <<- sum(bincounts)


# Calculate likelihood intervals in the untransformed space

        likelihood_min = rep(0,bins)

        likelihood_max = rep(0,bins)

        p = seq(0,1,0.0001)

        for (i in 1:bins) {

                if (bincounts[i]>0) {

                        loglike = dbinom(bincounts[i],N,prob=p,log=T)

                        interval = p[loglike>(max(loglike)-2)]

                        likelihood_min[i] = interval[1]

                        likelihood_max[i] = interval[length(interval)]

                        }

                }
```

```
# Transform density and likelihood intervals

binheights = rep(0,bins)

for (i in 1:bins) {

        if (bincounts[i]>0) {

                bin_end = sum(binwidths[1:i])

                bin_start = bin_end - binwidths[i]

                binheights[i] = (bincounts[i]/N)/(pi*(bin_end^2-bin_start^2))

                likelihood_min[i] = likelihood_min[i]/(pi*(bin_end^2-bin_start^2))

                likelihood_max[i] = likelihood_max[i]/(pi*(bin_end^2-bin_start^2))

                }

        }



trans = list()

trans$binheights = binheights

trans$likelihood_min = likelihood_min

trans$likelihood_max = likelihood_max

return(trans)

}
```